经济学前沿译丛

非参数计量经济学

理论与实践

〔加〕 李奇 (Qi Li)
〔美〕 杰弗里·斯科特·拉辛 (Jeffrey Scott Racine) 著
叶阿忠 吴相波 等译

Nonparametric Econometrics
Theory and Practice

Nonparametric Econometrics
Theory and Practice



非参数计量经济学
理论与实践

李奇和拉辛的《非参数计量经济学》对于任何一个认真地从事前沿问题研究的计量经济学家或统计学家来说都是一部必读的著作。本书同时覆盖了主流理论和相对冷僻的领域，因此对于非参数方法的理论介绍是相当全面的。我尤其欣赏作者对连续和离散回归量以及设定检验的处理，我从未见过以如此全面的方式对这一主题的处理。毫无疑问，我将以本书作为我的研究生计量经济学课程的教材，并用于我个人研究的参考。

——罗宾·西克尔斯，莱斯大学

很少有学生试图应用非参数方法来分析现实数据，这也许是因为缺少一本优秀的教科书来从直观上解释这类技术如何应用于现实，以及为什么要使用这类技术。而李奇和拉辛的这本书将同时服务于应用研究者和研究生。本书语言通俗易懂，任何具备基本计量经济学知识却对非参数计量方法毫不了解的人都可以理解它的内容。本书还包含了丰富的细节，从而清晰地阐述了实施这些方法的步骤。

——艾春荣，佛罗里达大学

本书对计量经济学领域做出了重要贡献。它全面地覆盖了非参数和半参数方法在经济模型和经济数据中的应用，而且第一次以一种简明的方式处理了与离散数据及混合数据相关的新内容。本书很好地平衡了理论与实践两个方面。除了可以作为研究生层次的优秀教材，我相信这本书也将成为很多研究人员必备的案头参考。

——大卫·E. 贾尔斯，维多利亚大学


"北京大学出版社"
微信公众号

上架建议：经济学
ISBN 978-7-301-24967-3

定价：79.00元

经济学前沿译丛

非参数计量经济学

理论与实践

〔加〕 李奇 (Qi Li)
〔美〕 杰弗里·斯科特·拉辛 (Jeffrey Scott Racine) 著
叶阿忠 吴相波 等译

cs Theory and Practice

北京大学出版社
PEKING UNIVERSITY PRESS

著作权合同登记号　图字:01－2009－3338

图书在版编目(CIP)数据

非参数计量经济学:理论与实践/(加)李奇,(美)拉辛著;叶阿忠等译.—北京:北京大学出版社,2015.1

(经济学前沿译丛)

ISBN 978－7－301－24967－3

Ⅰ.①非…　Ⅱ.①李…　②拉…　③叶…　Ⅲ.①计量经济学　Ⅳ.①F224.0

中国版本图书馆CIP数据核字(2014)第233738号

Nonparametric Econometrics: Theory and Practice

Qi Li and Jeffrey Scott Racine

书　　名	非参数计量经济学:理论与实践
著作责任者	〔加〕李奇(Qi Li)　〔美〕杰弗里·斯科特·拉辛(Jefferey Scott Racine)　著 叶阿忠　吴相波　等译
责任编辑	郝小楠
标准书号	ISBN 978－7－301－24967－3
出版发行	北京大学出版社
地　　址	北京市海淀区成府路205号　100871
网　　址	http://www.pup.cn
电子信箱	em@pup.cn　　QQ:552063295
新浪微博	@北京大学出版社　@北京大学出版社经管图书
电　　话	邮购部62752015　发行部62750672　编辑部62752926
印 刷 者	北京宏伟双华印刷有限公司
经 销 者	新华书店
	730毫米×1020毫米　16开本　36.5印张　655千字 2015年1月第1版　2015年1月第1次印刷
印　　数	0001—3000册
定　　价	79.00元

Nonparametric Econometrics

Theory and Practice

N o n p a r a m e t r i c   E c o n o m e t r

前　言

在本书中,术语“非参数”(nonparametric)是指不要求一个研究者对一个待估对象设定一个函数形式的统计技术。我们不是假定一个对象的函数形式是已知的,且取决于一些(有限的)未知参数,而是对研究对象用更少的限制性假设,如平滑性(可微性)和矩约束来替代。例如,当我们对估计一个地区的收入分布感兴趣时,不是假设其密度函数属于类似正态或对数正态族的一个参数族,而只是假设其密度函数是二阶(或三阶)可微的。当然,如果人们拥有关于感兴趣的对象的函数形式的先验知识(有人称之为“神圣的洞察力”(divine insight)),那么,他们通过使用参数技术将会做得更好。然而,在实践中,这些函数形式很少是已知的,并且不能忽视参数的错误设定所带来的众所周知的后果,这里不再赘述。 xvii

由于非参数技术对待估对象比参数技术做了更少的假设,非参数估计量倾向于比“正确设定”的参数估计量更慢地收敛于研究对象。此外,不像它们的参数对应部分,收敛速度通常跟包含的变量(协变量)数目反向相关,即变量越多收敛速度越慢,这有时被称作“维数诅咒”(curse of dimensionality)。然而,经常令人惊讶的是,即使对不是很大的数据集,非参数方法也能够揭示人们在使用普通参数函数形式时可能忽略的在这些数据中的结构。因此,非参数方法最适用于以下情况:(1) 人们对待估对象的函数形式所知甚少,(2) 变量(协变量)数目较小,(3) 研究者有一个相当大的数据集。第(2)和第(3)点是密切相关的,因为在非参数背景下,人们是否有一个足够大的样本取决于出现了多少协变量。Silverman(1986,见表4.2,p.94)对为得到准确的非参数估计而要求的样本大小和协变量个数之间的关系提供了一个非常好的说明。我们使用术语“半参数”(semipara- xviii

metric)来表示一种不要求研究者对待估对象的某一部分设定一个参数函数形式但要求对剩余部分进行参数假定的统计技术。

如上所述,相对于通常使用的参数模型,在本书中介绍的非参数方法具有对函数形式(如回归或条件概率函数)施加更少约束性假定的优点。然而,可供选择的方法可能通过放松在传统参数背景下的约束性假定而得到。Manski(2003)和他的合作者采用了这样一种方法来考虑一些参数不能被识别的概率或回归模型。对识别参数不是施加过于强的假设,对这些参数找到可行的区间边界通常是可能的。当边界相对较紧,也就是说,当可行的区间非常窄时,人们几乎可以识别这些参数。然而,这个令人兴奋的探索方向超出了本书的范畴,因此我们推荐感兴趣的读者参考 Manski(2003)的优秀专著;也可参见 Manski 和 Tamer(2002),Imbens 和 Manski(2004),Honoré 和 Tamer(2006)的著作以及其中的参考文献。

在过去的几十年间,非参数和半参数方法已经引起了许多统计学家的关注,这一点由统计学家所写的大量教材可以得到证实,其中包括 Prakasa Rao(1983),Devroye 和 Gyorfi(1985),Silverman(1986),Scott(1992),Bickel, Klaassen,Ritov 和 Wellner(1993),Wand 和 Jones(1995),Fan 和 Gijbels(1996),Simonoff(1996),Azzalini 和 Bowman(1997),Hart(1997),Efromovich(1999),Eubank(1999),Ruppert, Carroll 和 Wand(2003),以及 Fan 和 Yao(2005)。然而,专门适合应用计量经济学家需要的教材数目是相对稀少的,Härdle(1990),Horowitz(1998),Pagan 和 Ullah(1999),Yatchew(2003),以及 Härdle,Müller,Sperlich 和 Werwatz(2004)是我们目前已知的其中一些。

此外,大量已有的教材是从潜在数据在本质上是严格连续的这个假定开始,而在应用背景下经济学家更常处理的是分类(名义的和序数的)数据。用传统的基于频数的非参数方法来处理离散变量出现的情况被认为是不令人满意的。基于 Aitchison 和 Aitken(1976)关于平滑离散协变量的开创性工作,我们最近提出了许多新的非参数方法。例如,参见 Li 和 Racine(2003),Hall,Racine 和 Li(2004),Racine 和 Li(2004),Li 和 Racine(2004a),Racine,Li 和 Zhu(2004),Ouyang,Li 和 Racine(2006),Hall,Li 和 Racine(2006),Racine,Hart 和 Li(2006),Hsiao,Li 和 Racine(2007),以及 Li 和 Racine(2008)。在本书中,我们在一个统一的框架内强调适用于一个应用经济学家所能遇到的广泛的数据类型(连续、名义和序数)的非参数技术。

xix

本书的另一个重要特点是它强调在潜在不相关变量出现时非参数估计量的性质。已有的核方法处理,尤其是带宽选择方法假定所有变量是相关的。例如,已有的插入或交错鉴定处理方法假定在一个回归模型中所有协变量是相关

的，也就是说，所有协变量都有助于解释结果（即被解释变量）的变化。然而，当这种情况不成立时，已有的结果如收敛速度和带宽表现则不再成立。例如，参见 Hall 等（2004），Hall 等（2006），Racine 和 Li（2004），Li 和 Racine（2004a）。我们认为这是有效地进行非参数估计的一个非常重要的方面，如果应用者想灵活地运用这些工具就必须认识到这个事实。

本书是以选修了非参数和半参数方法的研究生课程的学生为对象，他们对经济学和其他社会科学等应用领域感兴趣。理想的先修课程可能包括一门数理统计的课程和一门参数计量经济学的课程，其层次相当于 Greene（2003）和 Wooldridge（2002）。我们也想让这本书作为参考服务于更为广泛的读者群，包括应用研究者和那些希望熟悉这个领域的人。

本书的五个部分组织如下：第 1 部分涵盖了独立数据的密度和回归函数的非参数估计，其重点被放在混合的离散和连续数据类型；第 2 部分再次处理了独立数据的各种半参数模型，包括局部线性模型、单一指数模型、可加模型、变系数模型审查模型以及样本选择模型；第 3 部分处理了一系列一致模型设定检验；第 4 部分研究了近邻和序列方法；第 5 部分考虑了工具变量模型、联立方程模型以及面板数据模型的核估计，并且把以前各章的结果扩展到弱依赖数据的背景中。

第 1 部分中的多数结果都给出了严格的证明，而其余部分的很多结果给出了证明的要点。作为背景知识的统计概念在附录中给出。

一个 R 软件包（R 开发核心团队（2006））可供使用而且可以直接从 http://www.R-project.org 得到，它能运行很多在第 1、第 2 部分讨论的方法以及一些在第 3、第 4、第 5 部分讨论的方法。它还包含了本书中使用的一些数据集，也包括了一个使读者很容易运行新的关于核检验和核估计量的函数。

习题出现在每章的末尾，并且为多数问题提供了详细的提示，鼓励想熟练掌握内容的学生做尽可能多的习题。因为一些提示可能几乎很简单地给出问题，我们强烈推荐想掌握技术的学生不要先参考提示再做题。

我们从提供指导、灵感或者使本书成为可能的奠定基础的许多人那里受益良多。全部列出他们来是不可能的。然而，我们想请每一个已经以某种方式为这本书做出贡献的人在它完成时与我们一起共同享受一种个人意义上的成就感。

在此，我们想感谢普林斯顿大学出版社的员工，他们的名字是 Peter Dougherty、Seth Ditchik、Terri O'Prey 和 Carole Schwager，感谢他们在这个过程中细致和专业的指导。

我们也想对许多基金机构表达我们深深的感激，感谢他们对资助构成本书

核心的研究的慷慨支持。李奇尤其想感谢来自加拿大社会科学和人类研究委员会(SSHRC)、加拿大自然科学和工程研究委员会(NSERC)、得克萨斯A&M大学私人研究中心以及布什学院经济学项目的支持。拉辛想感谢来自SSHRC、NSERC、雪城大学政策研究中心以及美国自然科学基金(NSF)的支持。

xxi

我们也想感谢在麦克马斯特大学、雪城大学、得克萨斯A&M大学、加利福尼亚大学圣迭戈分校,圭尔夫大学、南佛罗里达大学以及约克大学的研究生,他们作为无意识做出贡献的对象,对书稿的早期草稿提供了有价值的反馈。

我们想进一步对许多合作者的各种贡献表示感谢。我们将特别感谢Peter Hall,他关于混合数据类型的核方法,特别是他对有无关变量核方法的理论基础的单独贡献已经使这项工作取得了许多成果。

许多人提供的反馈已经深化了我们的理解并提高了这本书的水平。我们尤其要感谢艾春荣、蔡宗武、陈晓红、David Giles、范延琴、黄建华、孙义国和杨立坚。

在更个人的意义上,拉辛想表达他对Aman Ullah深深的感情和私人的亏欠,他不仅把他带入了非参数统计学的大门,还指导他的论文并且仍然是鼓励他继续前进的动力源泉。

最后,李奇想把这本书献给他的妻子刘振娟,他的女儿Kathy和儿子Kevin,没有他们的默默支持这个项目可能不会成为现实。李奇也想充满爱和感激地把这本书献给他的父母。拉辛想把这本书作为对他在2005年11月22日去世的父亲的深情纪念,他曾经是一个导向灯并将仍然是他永远的动力来源。拉辛也想把这本书献给他的妻子Jennifer和儿子Adam,他们超越他们的知识领域在继续不断地充实他的生命。

目　录

第1部分　非参数核方法

目 录

目 录

第2部分 半参数方法

目 录

目　录

目 录

目 录

目 录

目　录

第1部分　非参数核方法

第1章 密度估计

3 在社会科学领域，概率密度函数（PDF）和累积分布函数（CDF）的估计是数据分析的基础。检验两个分布（或其矩）的相同可能是所有数据分析中最基本的检验。例如，经济学家把大量的注意力用于研究收入分布以及它们在不同地区和时间如何变化。概率密度函数和累积分布函数不仅是人们通常直接关注的对象，而且它们的估计也是其他对象建模的重要基石，如条件均值（即“回归函数”），这些对象可直接使用非参数或半参数方法建模（条件均值是条件概率密度函数的函数，而条件概率密度函数本身是无条件概率密度函数的一个比值）。掌握概率密度函数非参数估计的基本原理后，在第2章介绍的数据分析的重要内容，即条件均值函数的非参数估计，将以一种相当自然的方式展开。认真学习第1章介绍的方法对理解后面章节出现的内容将是非常有帮助的。

我们在1.1节到1.3节以单变量概率密度函数的估计开始，1.4节和1.5节转向单变量累积分布函数的估计，接着在1.6节到1.8节介绍更一般的多变量情况。1.9节到1.12节介绍渐近正态性（asymptotic normality）、一致收敛速度（uniform rates of convergence）和偏误减少方法（bias reduction methods）。1.13节对这方面的应用做了许多阐述，而1.14节是一些理论和应用的习题。

我们现在讨论如何估计一个随机变量 X 的概率密度函数 $f_X(x)$。4 为了记号简单，我们省略下标 X 并只用 $f(x)$ 表示 X 的概率密度函数。在本章中讨论的概率密度函数核估计的一些处理参考了 Silverman（1986）和 Scott（1992）两本优秀著作。

1.1 单变量密度估计

为了更好地理解人们为什么会考虑使用非参数方法来估计概率密度函数,我们先从概率密度函数参数估计的一个说明性的例子开始。

例 1.1 假定 $X_1, X_2, \cdots, X_n$ 独立同分布(i. i. d.),取自一个均值为 μ 方差为 σ^2 的正态分布,我们打算估计这个正态分布的概率密度函数 $f(x)$。

根据假定,$f(x)$ 有一个已知的参数函数形式(如单变量正态分布),且 $f(x) = (2\pi\sigma^2)^{-1/2}\exp\left[-\frac{1}{2}(x-\mu)^2/\sigma^2\right]$,其中均值 $\mu = E(X)$ 和方差 $\sigma^2 = E[(X-E(X))^2] = \mathrm{var}(X)$ 为仅有的待估未知参数。我们可以通过下面的极大似然方法估计 μ 和 σ^2。在独立同分布(i. i. d.)的假定下,$X_1, X_2, \cdots, X_n$ 的联合概率密度函数是单变量概率密度函数的简单乘积,它可被写为

$$f(X_1, \cdots, X_n) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X_i-\mu)^2}{2\sigma^2}} = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2}$$

这是以观测到的样本为已知条件,取对数得到对数似然函数

$$\mathcal{L}(\mu, \sigma^2) \equiv \ln f(X_1, \cdots, X_n; \mu, \sigma^2)$$

$$= -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2$$

最大似然方法在给定的分布假设下,通过选择相应参数,使得我们手中观测到的样本出现的可能性最大。因此,似然函数(或者其一个单调变换,例如取对数)表示了给定观测到的样本下 μ 和 σ^2 取不同值的合理性。然后我们最大化关于这两个未知参数的似然函数。

5 对数似然函数最大化的一阶必要条件是 $\partial\mathcal{L}(\mu, \sigma^2)/\partial\mu = 0$ 和 $\partial\mathcal{L}(\mu, \sigma^2)/\partial\sigma^2 = 0$。解关于这两个未知参数 μ 和 σ^2 的一阶条件得

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad 和 \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\mu})^2$$

上面的 $\hat{\mu}$ 和 $\hat{\sigma}^2$ 分别是 μ 和 σ^2 的最大似然估计量,并且 $f(x)$ 的相应估计量是

$$\hat{f}(x) = \frac{1}{\sqrt{2\pi\hat{\sigma}^2}}\exp\left[-\frac{1}{2}\left(\frac{x-\hat{\mu}}{\hat{\sigma}}\right)^2\right]$$

当然,任何参数方法的"阿基里斯之踵"*就是在估计之前分析者必须为待估对象设定正确的参数函数形式。仔细想一想,参数方法在某种程度上是循环

* "阿基里斯之踵"(Achilles heel)来自希腊神话,意思是"唯一致命的弱点"。——译者注

论证的,因为我们最初准备估计一个未知的密度函数,但是却必须首先假定密度函数的形式实际上是已知的(当然,已知的函数形式里包含一些未知的参数)。由于已经把我们的估计置于密度函数形式已知的假定之上,因此我们自然会遇到参数模型被"误设"的可能性,即与抽样数据的总体不一致。例如,在上面的例子中假定 X 取自一个正态分布总体,我们实际上施加了许多潜在很强的约束性假定:对称(symmetry),单峰(uni-modality),单调递减远离峰值(monotonically decreasing away from the mode)等。如果真实的密度事实上是非对称或者拥有多峰,或是非单调远离峰值,那么这种分布正态性假定可能为真实的密度提供了一个误导性的特征,并且可能因此产生错误的估计和导致不合理的推断。

这时许多读者将很可能指出,由于已经估计了一个参数概率密度函数,人们能够检验这个分布的基本假定的有效性。当然,我们完全认可这种观点。然而,通常拒绝一个分布假定并不能够提供任何清晰的其他选择。也就是说,我们可以拒绝正态性假定,但是这种拒绝使得我们回到起点,只是排除了众多候选分布中的一个。基于这种背景,研究者可以考虑采用非参数方法替代。 6

非参数方法设法避开了在估计之前需要设定参数函数形式所产生的问题。它不是假定人们知道待估对象的准确函数形式,而是假定它满足一些规律条件,例如平滑(smoothness)和可微(differentiability)。然而,这不是没有代价的。通过对概率密度函数的函数形式施加比参数方法更少的结构,非参数方法需要更多的数据才能达到与正确设定的参数模型相同的精确度。在本书中我们的关注点是被称为"非参数核估计量"(nonparametric kernel estimators)的一系列估计量("核函数"只是一个加权函数),尽管在第 14 章和第 15 章我们提供了包括近邻和序列方法的不同的非参数方法。

在进行非参数密度估计方法的理论分析之前,我们首先考虑估计抛硬币时正面出现的概率这个常见例子,它跟一个累积分布函数的非参数估计紧密相关。这将依次把我们引入一个概率密度函数的非参数估计。

例 1.2　假定我们有一枚硬币(可能是不均匀的),我们想估计抛这枚硬币并且使得正面朝上的概率。令 $p=P(H)$ 表示(未知的)得到正面的总体概率。采用一个相对频率方法,我们将把这枚硬币抛 n 次,计算在 n 次试验中正面出现的频率,并且按照下式计算相对频率

$$\hat{p}=\frac{1}{n}\{\text{正面出现的次数}\} \tag{1.1}$$

这提供了 p 的一个估计。定义于式(1.1)的 $\hat{p}$ 经常被叫作 p 的"频率估计量"(frequency estimator),并且它也是 p 的最大似然估计量(见习题 1.2)。当然,估计量 $\hat{p}$ 是完全非参数的。直观地,人们将期望,如果 n 足够大,那么 $\hat{p}$ 应该"接

近"p。事实上,人们容易证明 $\hat{p}$ 的均方误差(MSE)由下式给出(见习题1.3)

$$\mathrm{MSE}(\hat{p}) \stackrel{\text{def}}{=} E[(\hat{p}-p)^2] = \frac{p(1-p)}{n}$$

因此当 $n\to\infty$ 时 $\mathrm{MSE}(\hat{p})\to 0$,这可被称作 $\hat{p}$ 在均方误差上收敛于 p;参见附录A对各种收敛形式的定义。

7 我们现在讨论如何得到 X 的累积分布函数的估计量,我们用 $F(x)$ 来表示。累积分布函数被定义为

$$F(x) = P[X \leqslant x]$$

$X_1,\cdots,X_n$ 是独立同分布的数据(即随机取自分布 $F(\cdot)$),人们能用下式估计 $F(x)$

$$F_n(x) = \frac{1}{n}\{X_i \leqslant x \text{ 的次数}\} \tag{1.2}$$

方程(1.2)有一个很好的直观解释。回到我们抛硬币的例子,如果硬币满足当我们抛它时得到正面的概率等于 $F(x)$($F(x)$未知),并且如果我们把数据集 $X_1,\cdots,X_n$ 看作把硬币抛 n 次,如果 $X_i\leqslant x$,我们说正面出现在第 i 次试验,那么 $P(H)=P(X_i\leqslant x)=F(x)$。我们所熟悉的 $P(H)$ 的频率估计量等于正面出现的次数除以试验次数:

$$\hat{P}(H) = \frac{\text{正面出现的次数}}{n} = \frac{1}{n}\{X_i \leqslant x \text{ 的次数}\} \equiv F_n(x) \tag{1.3}$$

因此,我们把式(1.2)叫作 $F(x)$ 的频率估计量。正如以前当估计 $P(H)$ 时,直觉上我们希望当 n 变大时,$\hat{P}(H)$ 应该是 $P(H)$ 的一个更准确的估计。类似的推理,人们将希望当 $n\to\infty$ 时,$F_n(x)$ 应该是 $F(x)$ 的一个更准确的估计。的确,人们能够很容易证明在均方误差上 $F_n(x)\to F(x)$,这意味着当 $n\to\infty$ 时 $F_n(x)$ 在概率和分布上收敛于 $F(x)$。在附录A中我们介绍了均方误差收敛、概率收敛、分布收敛和几乎确定收敛的概念。可以很容易地证明按这些不同收敛概念的任一种,$F_n(x)$ 均收敛于 $F(x)$。这些收敛的概念是必要的,因为很容易证明 $F_n(x)$ 的普通极限不存在,即 $\lim_{n\to\infty} F_n(x)$ 不存在(见习题1.3,其中给出了普通极限的定义)。这个例子说明了引入新的收敛模式概念的必要性,例如按均方误差收敛和按概率收敛。

现在我们考虑对函数形式不做参数假定时如何估计概率密度函数 $f(x)$ 这

8 个问题。从 $f(x)$ 的定义,我们有[①]

$$f(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(x) \tag{1.4}$$

① 在本章我们只考虑 X 是连续变量的情形。我们在第3章和第4章才处理 X 是离散的情况。

根据式(1.2)和式(1.4)，$f(x)$的一个显而易见的估计量是①

$$\hat{f}(x) = \frac{F_n(x+h) - F_n(x-h)}{2h} \tag{1.5}$$

其中h为一个小的正的增量。

把式(1.2)代入式(1.5)，我们得到

$$\hat{f}(x) = \frac{1}{2nh}\{X_1,\cdots,X_n\text{落入区间}[x-h,x+h]\text{的次数}\} \tag{1.6}$$

如果我们定义一个由下式给出的均匀核函数

$$k(z) = \begin{cases} 1/2, & |z| \leqslant 1 \\ 0, & \text{其他} \end{cases} \tag{1.7}$$

那么很容易看出由式(1.5)给出的$\hat{f}(x)$也可以被表示为

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{X_i - x}{h}\right) \tag{1.8}$$

等式(1.8)被叫作均匀核估计量，因为由式(1.7)定义的核函数$k(\cdot)$对应着均匀概率密度函数。一般地，我们把$k(\cdot)$称为核函数并且把h称为平滑参数(或者又称为带宽或窗宽)。式(1.8)有时被称作“朴素”(naive)核估计量。

事实上在这个背景下人们会对核函数$k(\cdot)$使用许多其他可能的选择。例如，人们可以使用下面给出的标准正态核

$$k(v) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}v^2}, \quad -\infty < v < \infty \tag{1.9}$$

这类估计量能够在由 Rosenblatt(1956)最早发表的关于核密度估计的论文中找到，而 Parzen(1962)证明了跟这类估计量有关的许多性质，并且为了得到更有效的估计量放松了非负性(nonnegativity)的假定。出于这个原因，这种方法有时被称作“Rosenblatt-Parzen 核密度估计”。

9

我们将简短地证明定义于式(1.8)的核估计量$\hat{f}(x)$是$f(x)$的一致估计量，只要其构成来自任何一般非负有界核函数$k(\cdot)$且满足

$$\begin{aligned} &\text{(i)} \int k(v)\,dv = 1 \\ &\text{(ii)}\ k(v) = k(-v) \\ &\text{(iii)} \int v^2 k(v)\,dv = \kappa_2 < \infty \end{aligned} \tag{1.10}$$

① 回忆函数$g(x)$导数的定义是$dg(x)/dx = \lim_{h\to 0}\frac{g(x+h)-g(x)}{h}$，或者等价地，$dg(x)/dx = \lim_{h\to 0}\frac{g(x+h)-g(x-h)}{2h}$。

注意对称条件(ii)意味着$\int vk(v)\mathrm{d}v = 0$。根据一致性,我们有在概率上$\hat{f}(x) \to f(x)$(在概率上收敛的定义见附录A)。注意定义于式(1.10)的$k(\cdot)$是(对称的)概率密度函数。至于最近有关非对称核的核方法研究,参见Abadir和Lawford(2004)。

为定义各种收敛模式,我们首先引入关于向量的"欧几里得范数"(欧几里得长度,Euclidean length)的概念。给定一个$q \times 1$向量$x = (x_1, x_2, \cdots, x_q)' \in \mathbb{R}^q$,我们使用$\| x \|$来表示$x$的欧几里得长度,它由下式定义

$$\| x \| = [x'x]^{1/2} \equiv \sqrt{x_1^2 + x_2^2 + \cdots + x_q^2}$$

当$q = 1$(标量)时,$\| x \|$就是x的绝对值。

在附录中我们讨论了符号$O(\cdot)$("大O")和$o(\cdot)$("小O")。令a_n为非随机序列。如果$|a_n| \leqslant Cn^{\alpha}$对所有足够大的$n$都成立,其中$\alpha$和$C(>0)$是常数,则我们说$a_n = o(n^{\alpha})$。类似地,如果当$n \to \infty$时,$a_n / n^{\alpha} \to 0$,则我们说$a_n = o(n^{\alpha})$。我们现在准备证明$\hat{f}(x)$具有均方误差一致性。

定理1.1 令$X_1, \cdots, X_n$表示某三阶可微的概率密度函数$f(x)$的独立同分布观测值,并令$f^{(s)}(x)$表示$f(x)$的第s阶导数$(s = 1, 2, 3)$。令x是在X支撑上的一个内点,并令$\hat{f}(x)$由式(1.8)定义。假定核函数$k(\cdot)$有界而且满足(1.10)。同时,当$n \to \infty$时,有$h \to 0$和$nh \to \infty$成立,那么

$$\begin{aligned} \mathrm{MSE}(\hat{f}(x)) &= \frac{h^4}{4}[\kappa_2 f^{(2)}(x)]^2 + \frac{\kappa f(x)}{nh} + o(h^4 + (nh)^{-1}) \\ &= O(h^4 + (nh)^{-1}) \end{aligned} \tag{1.11}$$

其中$\kappa_2 = \int v^2 k(v)\mathrm{d}v$,$\kappa = \int k^2(v)\mathrm{d}v$,并且对$l = 0, 1, 2, 3$,$\sup_{\xi \in S(X)} |f^{(l)}(\xi)| < \infty$,其中$S(X)$表示$X$的支撑,且$\int |v^3 k(v)| \mathrm{d}v < \infty$。

定理1.1的证明:

$$\begin{aligned} \mathrm{MSE}(\hat{f}(x)) &\equiv E\{[\hat{f}(x) - f(x)]^2\} \\ &= \mathrm{var}(\hat{f}(x)) + [E(\hat{f}(x)) - f(x)]^2 \\ &\equiv \mathrm{var}(\hat{f}(x)) + [\mathrm{bias}(\hat{f}(x))]^2 \end{aligned}$$

我们将分别展开偏误项$\mathrm{bias}(\hat{f}(x))$和方差项$\mathrm{var}(\hat{f}(x))$。

为计算偏误项,我们需要Taylor展开公式。对一个m阶可微的单变量函数$g(x)$,我们有

$$g(x) = g(x_0) + g^{(1)}(x_0)(x - x_0) + \frac{1}{2!}g^{(2)}(x_0)(x - x_0)^2 + \cdots$$

$$+\frac{1}{(m-1)!}g^{(m-1)}(x_0)(x-x_0)^{m-1}+\frac{1}{m!}g^{(m)}(\xi)(x-x_0)^m$$

其中 $g^{(s)}(x_0)=\left.\frac{\partial^s g(x)}{\partial x^s}\right|_{x=x_0}$,并且 ξ 位于 x 与 x_0 之间。

偏差项由下式给出: 11

$$\begin{aligned}\operatorname{bias}(\hat{f}(x)) &= E\left\{\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{X_i-x}{h}\right)\right\}-f(x)\\ &= h^{-1}E\left[k\left(\frac{X_1-x}{h}\right)\right]-f(x)\\ &\quad \text{(依同分布)}\\ &= h^{-1}\int f(x_1)k\left(\frac{x_1-x}{h}\right)\mathrm{d}x_1-f(x)\\ &= h^{-1}\int f(x+hv)k(v)h\mathrm{d}v-f(x)\\ &\quad \text{(变量变换, } x_1-x=hv\text{)}\\ &= \int\left\{f(x)+f^{(1)}(x)hv+\frac{1}{2}f^{(2)}(x)h^2v^2+O(h^3)\right\}k(v)\mathrm{d}v-f(x)\\ &= \left\{f(x)+0+\frac{h^2}{2}f^{(2)}(x)\int v^2k(v)\mathrm{d}v+O(h^3)\right\}-f(x)\\ &\quad \text{(由式(1.10))}\\ &= \frac{h^2}{2}f^{(2)}(x)\int v^2k(v)\mathrm{d}v+O(h^3)\end{aligned}\tag{1.12}$$

其中 $O(h^3)$ 项来自

$$(1/3!)h^3\left|\int f^{(3)}(\tilde{x})v^3k(v)\mathrm{d}v\right|\leqslant Ch^3\int|v^3k(v)|\mathrm{d}v=O(h^3)$$

其中 C 是一个正的常数,并且 $\tilde{x}$ 位于 x 与 $x+hv$ 之间。

注意在以上推导中我们假定 $f(x)$ 是三阶可微的。我们可以放松这个条件为 $f(x)$ 是二阶可微的,使得($O(h^3)$ 变为 $o(h^2)$,见习题 1.5)

$$\begin{aligned}\operatorname{bias}(\hat{f}(x)) &= E(\hat{f}(x))-f(x)\\ &= \frac{h^2}{2}f^{(2)}(x)\int v^2k(v)\mathrm{d}v+o(h^2)\end{aligned}\tag{1.13}$$

下一步我们考虑方差项,注意 12

$$\begin{aligned}\operatorname{var}(\hat{f}(x)) &= \operatorname{var}\left[\frac{1}{nh}\sum_{i=1}^{n}k\left(\frac{X_i-x}{h}\right)\right]\\ &= \frac{1}{n^2h^2}\left\{\sum_{i=1}^{n}\operatorname{var}\left[k\left(\frac{X_i-x}{h}\right)\right]+0\right\}\\ &\quad \text{(独立性)}\end{aligned}$$

$$= \frac{1}{nh^2}\mathrm{var}\left(k\left(\frac{X_1 - x}{h}\right)\right)$$

（同分布）

$$= \frac{1}{nh^2}\left\{E\left[k^2\left(\frac{X_1 - x}{h}\right)\right] - \left[E\left(k\left(\frac{X_1 - x}{h}\right)\right)\right]^2\right\}$$

$$= \frac{1}{nh^2}\left\{\int f(x_1)k^2\left(\frac{x_1 - x}{h}\right)\mathrm{d}x_1 - \left[\int f(x_1)k\left(\frac{x_1 - x}{h}\right)\mathrm{d}x_1\right]^2\right\}$$

$$= \frac{1}{nh^2}\left\{h\int f(x + hv)k^2(v)\,\mathrm{d}v - \left[h\int f(x + hv)k(v)\,\mathrm{d}v\right]^2\right\}$$

$$= \frac{1}{nh^2}\left\{h\int[f(x) + f^{(1)}(\xi)hv]k^2(v)\,\mathrm{d}v - O(h^2)\right\}$$

$$= \frac{1}{nh}\left\{f(x)\int k^2(v)\,\mathrm{d}v + O\left(h\int|v|k^2(v)\,\mathrm{d}v\right) - O(h)\right\}$$

$$= \frac{1}{nh}\{\kappa f(x) + O(h)\} \tag{1.14}$$

其中 $\kappa = \int k^2(v)\,\mathrm{d}v$。

式(1.12)和式(1.14)完成了定理 1.1 的证明。 □

定理 1.1 意味着(根据附录 A 中定理 A.7)

$$\hat{f}(x) - f(x) = O_p(h^2 + (nh)^{-1/2}) = o_p(1)$$

通过选择 $h = cn^{-1/\alpha}(c > 0, \alpha > 1)$，$f(x)$ 的估计是一致估计所需的条件 $h \to 0$ 和 $nh \to \infty$ 是明显满足的。最重要的问题是在应用中 c 和 α 应该使用什么值。正如所看到的，对一个给定的大小为 n 的样本，如果 h 很小，相应的估计量将有一个小的偏误，代价是方差较大。另一方面，如果 h 很大，那么相应的估计量将有一个小的方差，代价是偏误较大。为最小化 $\mathrm{MSE}(\hat{f}(x))$，人们应该平衡平方偏误和方差项。h 的最优选择(在最小化 $\mathrm{MSE}(\hat{f}(x))$ 的意义上)应该满足 $\mathrm{dMSE}(\hat{f}(x))/\mathrm{d}h = 0$。通过使用式(1.11)很容易证明，最小化 $\mathrm{MSE}(\hat{f}(x))$ 首项的最优 h 由下式给出

$$h_{\mathrm{opt}} = c(x)n^{-1/5} \tag{1.15}$$

其中 $c(x) = \{\kappa f(x)/[\kappa_2 f^{(2)}(x)]^2\}^{1/5}$。

$\mathrm{MSE}(\hat{f}(x))$ 很明显具有“逐点的”性质，并且通过使用这个性质作为窗宽选择的基础，当在点 x 处估计一个密度时我们会得到一个最优窗宽。考察式(1.15)中的 $c(x)$，我们会发现对位于一个分布尾部的点 x 的估计的最优窗宽将不同于对位于众数的点估计的最优窗宽。假定我们感兴趣的不是调整窗宽以适应 $f(x)$ 的逐点估计而是对所有的点 x，也就是对位于 $f(\cdot)$ 支撑($f(x)$ 的支撑被定义为满足

$f(x)>0$ 的点 x 的集合,即 $\{x:f(x)>0\}$)上的所有点 x,在整体上调整窗宽。在这种情况下,我们可以通过最小化 $\hat{f}(x)$ 的"积分均方误差"(IMSE)来选择最优的 h。使用式(1.11)我们有

$$\text{IMSE}(\hat{f}) \stackrel{\text{def}}{=} \int E[\hat{f}(x)-f(x)]^2 \mathrm{d}x = \frac{1}{4}h^4\kappa_2^2\int[f^{(2)}(x)]^2\mathrm{d}x + \frac{\kappa}{nh} + o(h^4+(nh)^{-1}) \tag{1.16}$$

再次令 h_{opt} 表示最小化式(1.16)首项的最优平滑参数,我们利用简单积分得到

$$h_{\text{opt}} = c_0 n^{-1/5} \tag{1.17}$$

其中 $c_0=\kappa_2^{-2/5}\kappa^{1/5}\left\{\int[f^{(2)}(x)]^2\mathrm{d}x\right\}^{-1/5}>0$ 是一个正常数。注意如果对(几乎)所有的 x 有 $f^{(2)}(x)=0$,那么,c_0 是没有很好定义的。例如,如果 X 均匀分布于它的支撑,那么,$f^{(s)}(x)=0$ 对所有 x 和所有 $s\geqslant1$ 成立,并且式(1.17)在这种情况下没有定义。可以证明,在这种情况(即当 X 是均匀分布时)下,h_{opt} 将有一个等于 $n^{-1/3}$ 的不同的收敛速度(参见1.3.1节和习题1.16的相关讨论)。　14

上述结果的一个有趣扩展可在 Zinde-Walsh(2005)中找到,该文通过广义函数和广义随机过程考察了核密度估计量的渐近性质,并提出了当密度不存在时,即作为局部加权函数,密度不存在时,刻画核密度估计量性质的新颖结果。

1.2　单变量窗宽选择:经验法则和插入法

方程(1.17)说明最优平滑参数通过 c_0 依赖于未知密度函数二阶导数的积分。在应用中,为了获得 $\int[f^{(2)}(x)]^2\mathrm{d}x$ 的非参数估计,人们可以选择 h 的一个初始"试验值"(pilot value),然后使用这个值通过式(1.17)得到 h_{opt}。这种方法顾名思义被称为"插入法"(plug-in methods)。Silverman(1986)提出一种流行的选择初始 h 的方法,就是假定 $f(x)$ 属于一个参数的分布族,然后使用式(1.17)计算 h。例如,如果 $f(x)$ 是方差为 σ^2 的正态概率密度函数,那么 $\int[f^{(2)}(x)]^2\mathrm{d}x=3/[8\pi^{1/2}\sigma^5]$。使用一个标准正态核,通过式(1.17),我们得到以下试验估计

$$h_{\text{pilot}} = (4\pi)^{-1/10}[(3/8)\pi^{-1/2}]^{-1/5}\sigma n^{-1/5} \approx 1.06\sigma n^{-1/5} \tag{1.18}$$

而后将其插入 $\int[\hat{f}^{(2)}(x)]^2\mathrm{d}x$,接着使用式(1.17)和以上结果得到 h_{opt}。插入法的一个不佳的性质是它不是完全自控的,因为人们需要选择一个 h 的初始值

来估计 $\int [f^{(2)}(x)]^2 dx$（参见 Marron，Jones 和 Sheather（1996）以及 Loader（1999）的进一步讨论）。

在大多数情况下，研究者将用式(1.18)本身作为窗宽选择。这就是所谓的"正态推断经验法则"法，由于它是对一个特定分布族的最优窗宽（在本情况中是正态族），如果潜在分布接近于一个正态分布，那么将会得到好的结果，并且出于探索的目的它当然有计算上的吸引力。在应用中，σ 被 $\{X_i\}_{i=1}^n$ 的样本标准差代替，而 Silverman(1986，p. 47)提议使用一个更稳健的分散程度测量指标，即

15 用 A 代替 σ，这一"适应性的"(adaptive)分散程度指标由下式给出：

$$A = \min(\text{标准差}, \text{四分位矩}/1.34)$$

为了选择适合样本数据的窗宽 h，我们现在把注意力转向讨论完全自动或者"数据驱动"(data-driven)的选择方法。

1.3 单变量窗宽选择：交错鉴定法

理论和实践都说明，非参数核估计量对核函数的选择相对不敏感。然而，非参数核估计量对窗宽选择是敏感的。由不同窗宽得到的潜在分布甚至具有完全不同的性质。如果核方法仅被用于"探索的"目的，那么人们可以先选择一个小的 h 值得到一个光滑程度不高的密度估计，并让眼睛凭直觉增加窗宽以获得光滑程度较好的密度估计。此外，人们可能选择 h 的一个取值范围并画出相应的密度估计。然而，出于严谨的分析和推断目的，必须采用被普遍认可的具有最优性质的窗宽选择方法。人们把窗宽选择的重要性类比于在函数级数展开中阶数的选择；在近似中包括的项越多，相应的模型就变得越灵活。而一个核估计量的窗宽越小，它就变得越灵活。然而，增加灵活性（减少潜在偏误）必然导致增加可变性（提高潜在方差）。从这方面来看，人们自然会考虑以下旨在平衡估计的偏误平方和方差的窗宽选择方法。

1.3.1 最小二乘交错鉴定

最小二乘交错鉴定(least square cross-validation)是一个选择平滑参数 h 的完全自控的数据驱动方法，最初由 Rudemo（1982）、Stone（1984）和 Bawman（1984）提出（也可参见 Silverman(1986，pp. 48—51)）。这种方法是基于这样一个选择窗宽的原则——最小化相应估计的积分平方误差，也就是说，它给出了适应于 $f(x)$ 的支撑中所有 x 的最优窗宽。

16 $\hat{f}$ 和 f 之差的平方的积分是

$$\int[\hat{f}(x)-f(x)]^2\mathrm{d}x=\int\hat{f}(x)^2\mathrm{d}x-2\int\hat{f}(x)f(x)\mathrm{d}x+\int f(x)^2\mathrm{d}x \quad (1.19)$$

由于式(1.19)右边第三项与 h 无关,因此选择 h 最小化式(1.19)等价于关于 h 最小化

$$\int\hat{f}(x)^2\mathrm{d}x-2\int\hat{f}(x)f(x)\mathrm{d}x \quad (1.20)$$

在第二项中,$\int\hat{f}(x)f(x)\mathrm{d}x$ 可被写为 $E_X[\hat{f}(X)]$,其中 $E_X(\cdot)$ 表示对 X 求期望而不是对用于计算 $\hat{f}(\cdot)$ 的随机观测值 $\{X_j\}_{j=1}^n$ 求期望。因此,我们可以用 $n^{-1}\sum_{i=1}^n\hat{f}_{-i}(X_i)$ 估计 $E_X[\hat{f}(X)]$(即用它的样本均值代替 E_X),其中

$$\hat{f}_{-i}(X_i)=\frac{1}{(n-1)h}\sum_{j=1,j\neq i}^n k\left(\frac{X_i-X_j}{h}\right) \quad (1.21)$$

是 $f(X_i)$ 的去一核估计量[①]。最后,我们通过下式估计首项 $\int\hat{f}(x)^2\mathrm{d}x$

$$\begin{aligned}\int\hat{f}(x)^2\mathrm{d}x&=\frac{1}{n^2h^2}\sum_{i=1}^n\sum_{j=1}^n\int k\left(\frac{X_i-x}{h}\right)k\left(\frac{X_j-x}{h}\right)\mathrm{d}x\\&=\frac{1}{n^2h}\sum_{i=1}^n\sum_{j=1}^n\bar{k}\left(\frac{X_i-X_j}{h}\right)\end{aligned} \quad (1.22)$$

其中 $\bar{k}(v)=\int k(u)k(v-u)\mathrm{d}u$ 是从 $k(\cdot)$ 得出的双重卷积核(twofold convolution kernel)。如果 $k(v)=\exp(-v^2/2)/\sqrt{2\pi}$,一个标准正态核,那么 $\bar{k}(v)=\exp(-v^2/4)/\sqrt{4\pi}$,一个均值为0和方差为2的正态核(即正态概率密度函数),这是由于两个独立随机变量 $N(0,1)$ 的和是随机变量 $N(0,2)$。

最小二乘交错鉴定方法选择 h 最小化

17

$$\begin{aligned}\mathrm{CV}_f(h)=&\frac{1}{n^2h}\sum_{i=1}^n\sum_{j=1}^n\bar{k}\left(\frac{X_i-X_j}{h}\right)\\&-\frac{2}{n(n-1)h}\sum_{i=1}^n\sum_{j\neq i,j=1}^n k\left(\frac{X_i-X_j}{h}\right)\end{aligned} \quad (1.23)$$

可通过典型的值搜寻算法获得窗宽的估计。

可以证明,$\mathrm{CV}_f(h)$ 的首项是由下式给出的 CV_{f0}(忽略与 h 不相关的那一项,见习题1.6)

---

① 这里我们强调使用去一核估计量(leave-one-out kernel estimator)来计算前面的 $E_X(\cdot)$ 是重要的。这是因为期望算子假定 X 和 $X_j(j=1,\cdots,n)$ 之间是相互独立的。如果不使用去一核估计量,将不能采用交错鉴定法,见习题1.6(iii)。

$$\mathrm{CV}_{f0}(h) = B_1 h^4 + \frac{\kappa}{nh} \tag{1.24}$$

其中 $B_1 = (\kappa_2^2/4)\left[\int[f^{(2)}(x)]^2\mathrm{d}x\right]\left(\kappa_2 = \int v^2 k(v)\mathrm{d}v, \kappa = \int k^2(v)\mathrm{d}v\right)$。因此,对(几乎)所有 x 只要 $f^{(2)}(x)$ 不为零,我们有 $B_1 > 0$。

令 h^0 表示最小化 CV_{f0}的 h 值。简单的微积分可以证明 $h^0 = c_0 n^{-1/5}$,其中

$$c_0 = [\kappa/(4B_1)]^{1/5} = \kappa^{1/5}\kappa_2^{-2/5}\left\{\int[f^{(2)}(x)]^2\mathrm{d}x\right\}^{-1/5}$$

比较 h^0 与在式(1.17)中的 h_{opt}表明这两者是相同的,即 $h^0 \equiv h_{\mathrm{opt}}$。这是因为 h_{opt} 最小化 $\int E[\hat{f}(x) - f(x)]^2\mathrm{d}x$,而 h^0 最小化 $E[\mathrm{CV}_f(h)]$,即 $\mathrm{CV}_f(h)$ 的首项。可以很容易看出 $E[\mathrm{CV}_f(h)] + \int f(x)^2\mathrm{d}x$ 是$\int E[\hat{f}(x) - f(x)]^2\mathrm{d}x$ 的另一种形式;因此,$E[\mathrm{CV}_f(h)] + \int f(x)^2\mathrm{d}x$ 也估计了$\int E[\hat{f}(x) - f(x)]^2\mathrm{d}x$。给定$\int f(x)^2\mathrm{d}x$ 与 h 不相关,人们将期望 h^0 与 h_{opt} 应该是一样的。

令 $\hat{h}$ 表示最小化 $\mathrm{CV}_f(h)$的 h 值。给定 $\mathrm{CV}_f(h) = \mathrm{CV}_{f0}(h) + (\mathrm{s.o.})$,其中(s. o.)表示(比 $\mathrm{CV}_{f0}(h)$)更低阶的项并且与 h 无关,可以证明 $\hat{h} = h^0 + o_p(h^0)$,或者,等价地有

$$\frac{\hat{h} - h^0}{h^0} \equiv \frac{\hat{h}}{h^0} - 1 \to 0 \text{ 依概率收敛} \tag{1.25}$$

直观地,式(1.25)很容易理解,因为 $\mathrm{CV}_f(h) = \mathrm{CV}_{f0}(h) + (\mathrm{s.o.})$,因此最小化 $\mathrm{CV}_f(h)$的 h 值应该渐近地接近于最小化 $\mathrm{CV}_{f0}(h)$的 h;因此,我们期望在式(1.25)的意义上 h^0 与 $\hat{h}$ 将彼此接近。Härdle, Hall 和 Marron(1988)证明 $(\hat{h} - h^0)/h^0 = O_p(n^{-1/10})$,它确实收敛于零(依概率),但收敛速度非常慢。

18

当构造如同在式(1.23)中出现的 CV_f 时,我们又一次强调了使用去一核估计量。如果人们使用标准核估计量来替代,最小二乘交错鉴定将不再成立,产生 $\hat{h} = 0$。习题 1.6 证明如果人们估计 $f(X_i)$时不使用去一核估计量,那么 $h = 0$ 最小化目标函数,这当然违反了当 $n \to \infty$ 时 $nh \to \infty$ 的一致性条件。

这里我们隐含地施加了 $f^{(2)}(x)$是非零函数的约束,这排除了 $f(x)$是均匀概率密度函数的情况。事实上这个条件可以放松。Stone(1984)证明,只要 $f(x)$是有界的,那么最小二乘交错鉴定法将从如下意义上选择最优的 h:

$$\frac{\int[\hat{f}(x,\hat{h}) - f(x)]^2\mathrm{d}x}{\inf_h \int[\hat{f}(x,h) - f(x)]^2\mathrm{d}x} \to 1 \quad \text{a. s.} \tag{1.26}$$

其中 $\hat{f}(x,\hat{h})$ 表示用交错鉴定法选择 $\hat{h}$ 的 $f(x)$的核估计量,并且 $\hat{f}(x,h)$是对一

般 h 的核估计量。显然，定义于式(1.26)的比率对任何 n 应该大于或等于 1。因此，Stone(1984)的结果说明，交错鉴定平滑参数选择在最小化积分平方误差估计的意义上是渐近最优的。在习题 1.16 中，我们进一步讨论即使当 $f(x)$ 是均匀概率密度函数时，为什么 $\hat{h}\to 0$ 的直观背景。

1.3.2　似然交错鉴定

似然交错鉴定(likelihood cross-validation)是另一种选择平滑参数 h 的自控数据驱动方法。由于这种估计在 Kullback-Leibler 意义上将接近于真实密度，因此，这种方法产生一个具有熵理论解释的密度估计。这种方法是由 Duin(1976)提出的。

似然交错鉴定选择 h 来最大化下式给出的(去一)对数似然函数

$$\mathcal{L} = \ln L = \sum_{i=1}^{n} \ln \hat{f}_{-i}(X_i)$$

其中 $\hat{f}_{-i}(X_i)$ 是在式(1.21)中定义的 $f(X_i)$ 的去一核估计量。似然交错鉴定的主要问题是它受 $f(x)$ 尾部行为的严重影响，并且当使用普通核函数时(见 Hall(1987a,1987b))对厚尾分布会引起不一致的结果。出于这个原因，似然交错鉴定法在统计文献中已经很少引起注意。

19

然而，似然交错鉴定法可能对某类标准分布的估计效果会很好(即瘦尾)。在 1.3.3 节当我们比较不同窗宽选择方法对相应密度估计的影响时，我们考虑似然交错鉴定的表现，并且在 1.13 节中我们考虑实证应用。

1.3.3　数据驱动窗宽选择的一个实例

图 1.1 给出了根据取自一个模拟的双峰分布($n=500$ 观测值)构造的核估计。二阶高斯(正态)核始终被使用，并且最小二乘交错鉴定窗宽选择的估计结果见图形左上部曲线，其中 $h_{lscv}=0.19$。图形右上部曲线也画出了基于正态推断经验法则($h_{ref}=0.34$)的估计，图形左下部曲线为一个低度平滑估计(窗宽为 $1/5\times h_{lscv}$)，图形右下部曲线为一个过度平滑估计(窗宽为 $5\times h_{lscv}$)。[①]

图 1.1 说明这个数据集的最小二乘交错鉴定看来产生了一个合理的密度估计，而经验法则推断是不合适的，因为它在某种程度上过度平滑。极端过度平滑可能产生一个单峰估计，它完全模糊了这个潜在分布的真实双峰特征。同时，低度平滑引起太多错误的峰。参见习题 1.17 的一个经验应用，它研究了低度和过度平滑对相应密度估计的影响。

① 似然交错鉴定法产生一个 $h_{mlcv}=0.15$ 的窗宽，对这个数据集这导致一个跟基于最小二乘交错鉴定法相同的密度估计。

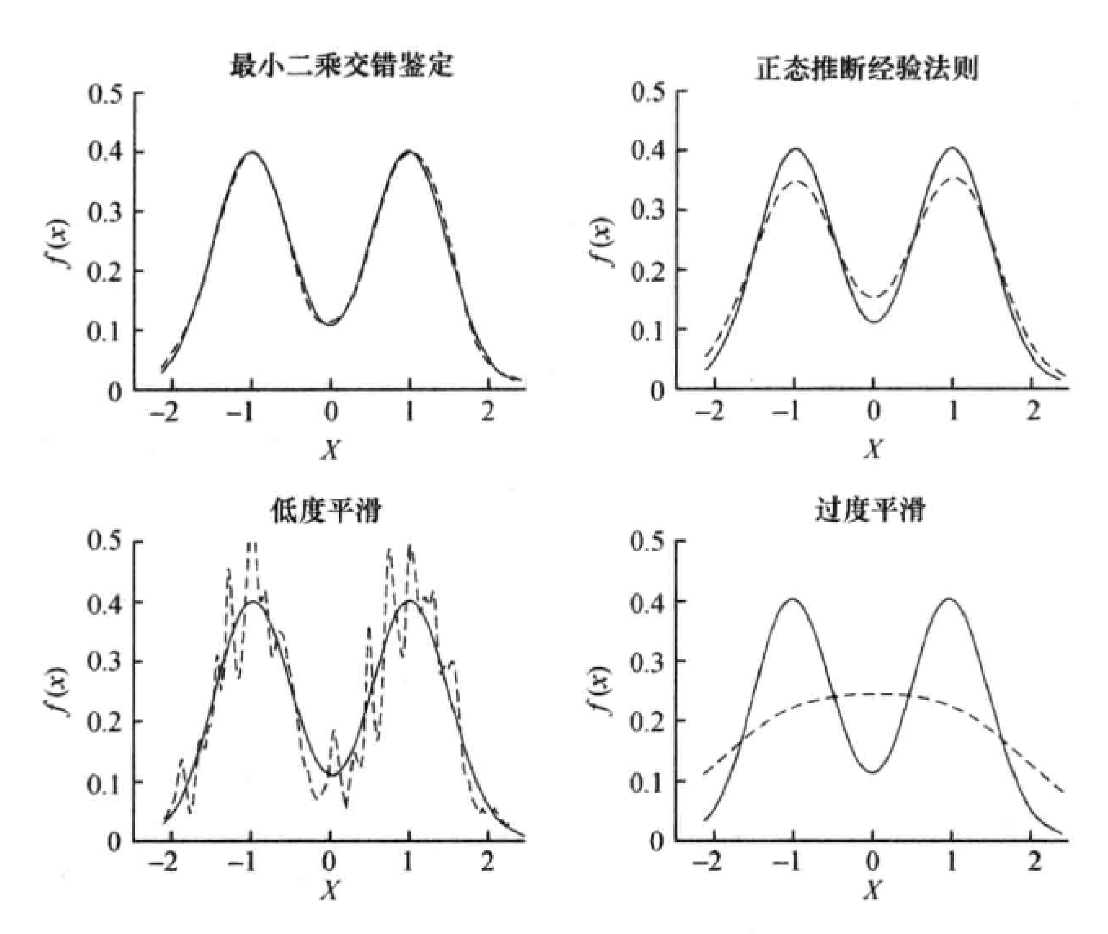


图 1.1　使用最小二乘交错鉴定、正态推断经验法则、低度平滑和过度平滑($n=500$)的混合正态的单变量核估计

注:实线是实际的密度函数,虚线是核估计。

1.4　单变量累积分布函数估计

在 1.1 节我们介绍了在式(1.2)中给出的经验累积分布函数估计量 $F_n(x)$,而习题 1.4 证明它是 $F(x)$ 的一个$\sqrt{n}$一致估计量。然而,这种经验的累积分布函数 $F_n(x)$ 是不平滑的,因为它在每一个样本实现点以 $1/n$ 跳跃。然而,人们通过积分 $\hat{f}(x)$ 能得到一个 $F(x)$ 的平滑估计。定义

$$\hat{F}(x)=\int_{-\infty}^{x}\hat{f}(v)\,\mathrm{d}v=\frac{1}{n}\sum_{i=1}^{n}G\left(\frac{x-X_i}{h}\right)\tag{1.27}$$

其中 $G(x)=\int_{-\infty}^{x}k(v)\,\mathrm{d}v$ 是一个累积分布函数(因为 $k(\cdot)$ 是一个概率密度函数,见式(1.10))。下一个定理给出了 $\hat{F}(x)$ 的均方误差。

定理 1.2　在 Bowman, Hall 和 Prvan(1998)给出的条件下,特别地,假定 $F(x)$ 是二阶连续可微的,$k(v)=\mathrm{d}G(v)/\mathrm{d}v$ 是有界、对称和紧支撑的,并且 $\mathrm{d}^2F(x)/\mathrm{d}x^2$ 是 Hölder 连续的,对某些 $0<\varepsilon<\frac{1}{8}$,有 $0\leqslant h\leqslant Cn^{-\varepsilon}$,那么当

$n \to \infty$ 时,

$$\begin{aligned} \mathrm{MSE}(\hat{F}) &= E[\hat{F}(x) - F(x)]^2 \\ &= c_0(x)n^{-1} - c_1(x)hn^{-1} + c_2(x)h^4 + o(h^4 + hn^{-1}) \end{aligned}$$

其中 $c_0(x) = F(x)(1 - F(x))$, $c_1(x) = \alpha_0 f(x)$, $\alpha_0 = 2\int vG(v)k(v)\mathrm{d}v$, $f(x) = \mathrm{d}F(x)/\mathrm{d}x$, $c_2(x) = [(\kappa_2/2)F^{(2)}(x)]^2$, $\kappa_2 = \int v^2 k(v)\mathrm{d}v$, 并且其中 $F^{(s)}(x) = \mathrm{d}^s F(x)/\mathrm{d}x^s$ 是 $F(x)$ 的第 s 阶导数。

证明:注意 $E[\hat{F}(x)] = E\left[G\left(\frac{x - X_i}{h}\right)\right]$。那么我们有 $\left(\int = \int_{-\infty}^{\infty}\right)$

$$\begin{aligned} E\left[G\left(\frac{x - X_i}{h}\right)\right] &= \int G\left(\frac{x - z}{h}\right)f(z)\mathrm{d}z \\ &= h\int G(v)f(x - hv)\mathrm{d}v = -\int G(v)\mathrm{d}F(x - hv) \\ &= -[G(v)F(x - hv)]\,|_{v=-\infty}^{v=\infty} + \int k(v)F(x - hv)\mathrm{d}v \\ &= \int k(v)[F(x) - F^{(1)}(x)hv + (1/2)h^2 F^{(2)}(x)v^2]\mathrm{d}v + o(h^2) \\ &= F(x) + (1/2)\kappa_2 h^2 F^{(2)}(x) + o(h^2) \end{aligned} \tag{1.28}$$

其中在上面第二个等式我们使用

$$-\int_{\infty}^{-\infty}[\cdots]\mathrm{d}v = \int_{-\infty}^{\infty}[\cdots]\mathrm{d}v$$

同时注意我们在 $\int G(v)f(x - hv)\mathrm{d}v$ 没有使用 Taylor 展开,因为对任何 $m \geq 0$ 有 $\int v^m G(v)\mathrm{d}v = +\infty$。我们首先进行分部积分, 然后使用 Taylor 展开, 因为 $\int v^m k(v)\mathrm{d}v$ 通常是有限的。例如,如果 $k(v)$ 是有界支撑或者 $k(v)$ 是标准正态核函数,那么对任何 $m \geq 0$, $\int v^m k(v)\mathrm{d}v$ 是有限的。

类似地,　　22

$$\begin{aligned} E\left[G^2\left(\frac{x - X_i}{h}\right)\right] &= \int G^2\left(\frac{x - z}{h}\right)f(z)\mathrm{d}z = h\int G^2(v)f(x - hv)\mathrm{d}v \\ &= -\int G^2(v)\mathrm{d}F(x - hv) \\ &= 2\int G(v)k(v)F(x - hv)\mathrm{d}v \\ &= 2\int G(v)k(v)[F(x) - F^{(1)}(x)hv]\mathrm{d}v + O(h^2) \end{aligned}$$

$$= F(x) - \alpha_0 hf(x) + O(h^2) \tag{1.29}$$

其中 $\alpha_0 = 2\int vG(v)k(v)\mathrm{d}v$, 并且其中我们已经使用了如下事实

$$2\int_{-\infty}^{\infty} G(v)k(v)\mathrm{d}v = \int_{-\infty}^{\infty} \mathrm{d}G^2(v) = G^2(\infty) - G^2(-\infty) = 1$$

因为 $G(\cdot)$是一个(用户设定的)累积分布函数核函数。

从式(1.28)中我们有偏误 $\text{bias}[\hat{F}(x)] = (1/2)\kappa_2 h^2 F^{(2)}(x) + o(h^2)$,并且从式(1.28)和式(1.29)我们有

$$\begin{aligned}
\text{var}[\hat{F}(x)] &= n^{-1}\text{var}\left[G\left(\frac{x - X_i}{h}\right)\right] \\
&= n^{-1}\left\{E\left[G^2\left(\frac{x - X_i}{h}\right)\right] - \left[EG\left(\frac{x - X_i}{h}\right)\right]^2\right\} \\
&= n^{-1}F(x)[1 - F(x)] - \alpha_0 f(x)hn^{-1} + o(h/n)
\end{aligned}$$

因此,

$$\begin{aligned}
E(\hat{F}(x) - F(x))^2 &= [\text{bias}(\hat{F}(x))]^2 + \text{var}[\hat{F}(x)] \\
&= n^{-1}F(x)[1 - F(x)] + h^4(\kappa_2/2)^2[F^{(2)}(x)]^2 \\
&\quad - \alpha_0 f(x)\frac{h}{n} + o(h^4 + n^{-1}h)
\end{aligned} \tag{1.30}$$

这完成了定理1.2的证明。 □

23

从定理1.2我们立刻得到关于 $\hat{F}$ 的积分均方误差的如下结果:

$$\begin{aligned}
\text{IMSE}(\hat{F}) &= \int E[\hat{F}(x) - F(x)]^2\mathrm{d}x \\
&= C_0 n^{-1} - C_1 hn^{-1} + C_2 h^4 + o(h^4 + hn^{-1})
\end{aligned} \tag{1.31}$$

其中 $C_j = \int c_j(x)\mathrm{d}x (j = 0,1,2)$。令 h_0 表示最小化积分均方误差首项的 h 值,我们得到

$$h_0 = a_0 n^{-1/3}$$

其中 $a_0 = [C_1/(4C_2)]^{1/3}$,因此估计单变量累积分布函数的最优平滑参数比估计一个单变量概率密度函数有一个更快的收敛速度($n^{-1/3}$ 与 $n^{-1/5}$)。由 $h \sim n^{-1/3}$,我们有 $h^2 = O(n^{-2/3}) = o(n^{-1/2})$。因此,根据 Liapunov 中心极限定理,在分布上$\sqrt{n}[\hat{F}(x) - F(x)] \to N(0,F(x)[1 - F(x)])$;见附录 A 中引理 A.5 和其他一些有用的中心极限定理。

如同非参数概率密度函数估计一样,非参数累积分布函数的估计有广泛的潜在应用,虽然它并未被广泛使用。例如,它可被用于对没有施加参数假定的潜在累积分布函数检验随机占优,可参考 Barrett 和 Donald(2003)以及 Linton, Whang 和 Maasoumi(2005)。

1.5 单变量累积分布函数窗宽选择:交错鉴定法

Bowman 等(1998)建议通过最小化如下的交错鉴定函数为 $\hat{F}(x)$ 选择 h:

$$\mathrm{CV}_F(h) = \frac{1}{n}\sum_{i=1}^{n}\int\{\mathbf{1}(X_i \leqslant x) - \hat{F}_{-i}(x)\}^2 \mathrm{d}x \tag{1.32}$$

其中 $\hat{F}_{-i}(x) = (n-1)^{-1}\sum_{j\neq i}^{n} G\left(\frac{x - X_j}{h}\right)$是 $F(x)$ 的去一估计量。

Bowman 等(1998)证明 $\mathrm{CV}_F = E[\mathrm{CV}_F] + (\mathrm{s.o.})$并且有(见习题1.9)　　24

$$E[\mathrm{CV}_F(h)] = \int F(1-F)\mathrm{d}x + \frac{1}{n-1}\int F(1-F)\mathrm{d}x - C_1 hn^{-1} + C_2 h^4 + o(hn^{-1} + h^4) \tag{1.33}$$

我们观察到式(1.33)与式(1.31)中给出的 $\mathrm{IMSE}(\hat{F})$有相同的首项。因此,在渐近性质上,通过交错鉴定选择 h 为 $\hat{F}(x)$带来跟使用最优确定性平滑参数 h_0 时可能产生的同样的渐近最优性质。如果我们令 $\hat{h}$ 表示交错鉴定平滑参数,那么可以证明在概率上 $\hat{h}/h_0 \to 1$。注意当使用 $\hat{h}$ 时,$\hat{F}(x,\hat{h})$的渐近分布与 $\hat{F}(x,h_0)$的一样(通过使用附录A中给出的一个随机等连续性论证),即

$$\sqrt{n}(\hat{F}(x) - F(x)) \xrightarrow{d} N(0, F(x)(1 - F(x))) \tag{1.34}$$

其中 $\hat{F}(x)$定义于式(1.27),但用 $\hat{h}$ 取代了 h。注意在式(1.34)中没有偏误项出现,因为 $\mathrm{bias}(\hat{F}(x)) = O(h_0^2) = O(n^{-2/3}) = o(n^{-1/2})$,对概率密度函数估计不是这种情形。这里平方误差项有比主要方差项 $O(n^{-1})$(即 $\mathrm{var}(\hat{F}(x)) = O(n^{-1})$)更小的阶。

现在把我们的注意力转到以上提出的单变量核估计量的扩展,也就是多元核估计量上。我们在这一章仅考虑连续变量的情况;我们在第3章和第4章处理离散数据与连续和离散数据混合的情形。

1.6 多变量密度估计

假定 $X_1,\cdots,X_n$ 构成一个独立同分布且有一个共同概率密度函数 $f(x) = f(x_1,x_2,\cdots,x_q)$的 q 维向量($X_i \in \mathbb{R}^q$,某些 $q>1$)。令 X_{is}表示 X_i 的第 $s(s=1,\cdots,q)$个元素。使用一个从单变量核函数的乘积构造的"乘积核函数",我们用下式估计这个概率密度函数 $f(x)$:

$$\hat{f}(x) = \frac{1}{nh_1\cdots h_q}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) \tag{1.35}$$

25 其中$K\left(\frac{X_i-x}{h}\right)=k\left(\frac{X_{i1}-x_1}{h_1}\right)\times\cdots\times k\left(\frac{X_{iq}-x_q}{h_q}\right)$,并且其中$k(\cdot)$是一个满足式(1.10)的单变量核函数。

$\hat{f}(x)$的均方误差一致性的证明类似于单变量的情况。特别是,可以证明

$$\text{bias}(\hat{f}(x))=\frac{\kappa_2}{2}\sum_{s=1}^{q}h_s^2 f_{ss}(x)+O\left(\sum_{s=1}^{q}h_s^3\right) \tag{1.36}$$

其中$f_{ss}(x)$是$f(x)$对x_s的二阶导数,$\kappa_2=\int v^2k(v)\mathrm{d}v$,并且人们也能证明

$$\text{var}(\hat{f}(x))=\frac{1}{nh_1\cdots h_q}\left[\kappa^q f(x)+O\left(\sum_{s=1}^{q}\right)h_s^2\right]=O\left(\frac{1}{nh_1\cdots h_q}\right) \tag{1.37}$$

其中$\kappa=\int k^2(v)\mathrm{d}v$。式(1.36)和式(1.37)的证明与单变量$X$的情况类似,留作习题(见习题1.11)。

总之,我们得到如下结果

$$\begin{aligned}\text{MSE}(\hat{f}(x))&=[\text{bias}(\hat{f}(x))]^2+\text{var}(\hat{f}(x))\\&=O\left(\left(\sum_{s=1}^{q}h_s^2\right)^2+(nh_1\cdots h_q)^{-1}\right)\end{aligned}$$

因此,如果当$n\to\infty$时,$\max_{1\leqslant s\leqslant q}h_s\to 0$并且$nh_1\cdots h_q\to\infty$,那么我们在MSE上有$\hat{f}(x)\to f(x)$(依均方误差收敛),这意味着在概率上$\hat{f}(x)\to f(x)$(依概率收敛)。

如同我们在单变量情形中见到的,最优平滑参数h_s应该权衡平方偏误和方差项,即对所有s有$h_s^4=O((nh_1\cdots h_q)^{-1})$。因此,对某些正常数$c_s(s=1,\cdots,q)$,我们有$h_s=c_s n^{-1/(q+4)}$。在1.3节讨论的交错鉴定法能够很容易被扩展到多变量数据情形中,并且我们能够证明最小二乘交叉鉴定能够在1.3节提出的意义上选择最优的h_s(见下面1.8节)。

我们对上面的证明中需要的独立性假定进行简单评论。我们的假定是,在不同i之间数据是独立的。注意对每一个分量$X_{is}(s=1,\cdots,q)$没有限制条件施加于下标s。乘积核仅仅出于方便被使用,并且当然不要求X_{is}在下标s之间是26 独立的。换句话说,多元核密度估计量(1.35)是能够捕获X_i的不同分量之间的一般依赖性的。进一步,我们在18章中将放松这个“观测值之间独立”的假定,并且将看到上面所得到的所有结果适用于弱依赖数据设定。

1.7 多变量窗宽选择:经验法则和插入法

在1.2节我们在一个单变量背景下讨论了正态推断经验法则和插入法的使用。这个单变量正态推断经验法则可直接扩展为多变量情形。令q为X_i的

维数,可以对 $s=1,\cdots,q$ 选择 $h_s=c_sX_{s,sd}n^{-1/(4+q)}$,其中 $X_{s,sd}$ 是 $\{X_{is}\}_{i=1}^{n}$ 的样本标准差,且 c_s 是一个正常数。在实践中,人们仍然会遇到如何选择 c_s 的问题。对所有的 $s=1,\cdots,q$ 选择 $c_s=1.06$ 是有计算上的优势的;然而,这种选择把不同的 X_{is}对称地对待。在实践中,可能联合概率密度函数在一个维度(比如在 x_1)变化快,但在另一个维度(比如在 x_2)变化慢,那么人们应该选择一个相对较小的 c_1 值(因此一个小的 h_1)和一个相对较大的 c_2 值(因此一个大的 h_2)。不像我们即将讨论的交错鉴定法,经验法则没有提供这种灵活性。

另一方面,对插入法,$\hat{f}(x)$ 的(平方)偏误首项和方差首项必须被估计,然后必须选择 $h_1\cdots h_q$ 以最小化 $\hat{f}(x)$ 的均方误差首项。然而,$\hat{f}(x)$ 的均方误差首项包括未知的 $f(x)$ 和它的偏导函数,以及为估计这些未知的函数必须为每个变量选择的试验窗宽。如何最好地选择初始的试验平滑参数在高维情形中可能是较复杂的,并且据我们所知,在实际应用中插入法没有被广泛地使用,我们除了用于数据分析探索外也不建议使用它们。

1.8　多变量窗宽选择:交错鉴定法

27

1.8.1　最小二乘交错鉴定

在 1.3.1 节讨论的单变量最小二乘交错鉴定法可以被很好地扩展到多变量密度估计情形中。用一个多变量乘积核取代式(1.23)中的单变量核函数,交错鉴定法的目标函数为

$$\begin{aligned}\mathrm{CV}_f(h_1,\cdots,h_q) &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{K}_h(X_i,X_j)\\&\quad-\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i,j=1}^{n}K_h(X_i,X_j)\end{aligned} \tag{1.38}$$

其中

$$K_h(X_i,X_j) = \prod_{s=1}^{q}h_s^{-1}k\left(\frac{X_{is}-X_{js}}{h_s}\right)$$

$$\bar{K}_h(X_i,X_j) = \prod_{s=1}^{q}h_s^{-1}\bar{k}\left(\frac{X_{is}-X_{js}}{h_s}\right)$$

并且 $\bar{k}(v)$ 是基于 $k(\cdot)$ 的双重卷积核(twofold convolution kernel),其中 $k(\cdot)$ 是一个满足式(1.10)的单变量核函数。

习题 1.12 表明,$\mathrm{CV}_f(h_1,\cdots,h_q)$ 的首项由下式给出(忽略与 h_s 无关的项)

$$\mathrm{CV}_{f0}(h_1,\cdots,h_q) = \int\left[\sum_{s=1}^{q}B_s(x)h_s^2\right]^2\mathrm{d}x+\frac{\kappa^q}{nh_1\cdots h_q} \tag{1.39}$$

其中 $B_s(x)=(\kappa_2/2)f_{ss}(x)$。

通过 $h_s=a_s n^{-1/(q+4)}(s=1,\cdots,q)$ 定义 a_s，我们有

$$\mathrm{CV}_{f0}(h_1,\cdots,h_q) = n^{-4/(q+4)}\chi_f(\alpha_1,\cdots,\alpha_q) \tag{1.40}$$

其中

$$\chi_f(a_1,\cdots,a_q) = \int\left[\sum_{s=1}^{q} B_s(x)a_s^2\right]^2 \mathrm{d}x + \frac{\kappa^q}{a_1\cdots a_q} \tag{1.41}$$

28

令 a_s^0 是 a_s 最小化 $\chi_f=(a_1,\cdots,a_q)$ 的值。在和用于单变量情形的同样条件下，再假定 $f_{ss}(x)$ 对所有的 s 不是零函数，Li 和 Zhou(2005)证明每一个 a_s^0 是被唯一定义的、正的，并且有限(参见习题1.10)。令 $h_1^0,\cdots,h_q^0$ 表示最小化 CV_{f0} 的 $h_1,\cdots,h_q$ 值。那么从式(1.40)中我们知道 $h_s^0=a_s^0 n^{-1/(q+4)}=O(n^{-1/(q+4)})$。

习题1.12表明 CV_{f0} 也是 $E[\mathrm{CV}_f]$ 的首项。因此，非随机平滑参数 h_s^0 可以被解释为使得积分均方误差首项最小化的最优平滑参数。

令 $\hat{h}_1,\cdots,\hat{h}_q$ 表示最小化 CV_f 的 $h_1,\cdots,h_q$ 值。利用 $\mathrm{CV}_f=\mathrm{CV}_{f0}+(\mathrm{s.o.})$ 这个事实，我们能够证明 $\hat{h}_s=h_s^0+o_p(h_s^0)$。于是，我们有

$$\frac{\hat{h}_s-h_s^0}{h_s^0}=\frac{\hat{h}_s}{h_s^0}-1\to 0 \quad \text{依概率收敛}, s=1,\cdots,q \tag{1.42}$$

因此，通过交错鉴定法选择的平滑参数跟非随机最优平滑参数有相同的渐近最优性质。

注意如果对某些 s 几乎处处 $f_{ss}(x)=0$，那么 $B_s=0$ 并且以上结果不再成立。Stone(1984)证明交错鉴定法仍然在最小化积分平方误差估计的意义上最优地选择了 $h_1,\cdots,h_q$；也可参见 Ouyang 等(2006)对这种情形的一个更详细的讨论。

1.8.2 似然交错鉴定

多变量模型似然交错鉴定法直接通过最大化在1.3.2节给出的似然函数(多变量)得到，因此我们这里不再深入讨论。然而，我们需要指出，尽管执行很简单，它也遇到跟单变量情况出现厚尾分布时同样的缺陷(即它在这种情况下有过度平滑的倾向)。

1.9 密度估计量的渐近正态性

在本节我们证明 $\hat{f}(x)$ 有一个渐近正态分布。最流行的中心极限定理是在

29

附录A的引理A.3给出的 Lindeberg-Levy 中心极限定理，它指出在分布上 $n^{1/2}\left[n^{-1}\sum_{i=1}^{n}Z_i\right]\to N(0,\sigma^2)$，其中 Z_i 是独立同分布，均值为0，方差为 σ^2。尽管

Lindeberg-Levy 中心极限定理能够被用于推导第7、第8和第9章讨论的各种半参数估计量的渐近分布,但它不能被用于推导$\hat{f}(x)$ 的渐近分布。这是因为$\hat{f}(x) = n^{-1}\sum_i Z_{i,n}$,其中被加数$Z_{i,n} = K_h(X_i, x)$依赖于$n$(因为$h = h(n)$)。我们将使用附录A中引理A.5给出的Liapunov中心极限定理。

定理1.3 令$X_1, \cdots, X_n$是独立同分布的q维向量,其概率密度函数$f(\cdot)$是三阶连续可导且有界的。令x是X支撑的一个内点。如果当$n \to \infty$, $h_s \to 0$时,对所有的$s = 1, \cdots, q$,都有$nh_1 \cdots h_q \to \infty$,并且$(nh_1 \cdots h_q)\sum_{s=1}^{q} h_s^6 \to 0$,那么

$$\sqrt{nh_1 \cdots h_q}\left[\hat{f}(x) - f(x) - \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2 f_{ss}(x)\right] \xrightarrow{d} N(0, \kappa^q f(x))$$

证明:使用式(1.36)和式(1.37),容易证明

$$\sqrt{nh_1 \cdots h_q}\left[\hat{f}(x) - f(x) - \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2 f_{ss}(x)\right]$$

有渐近均值0和渐近方差$\kappa^q f(x)$,即

$$\begin{aligned}
&\sqrt{nh_1 \cdots h_q}\left[\hat{f}(x) - f(x) - \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2 f_{ss}(x)\right] \\
&= \sqrt{nh_1 \cdots h_q}[\hat{f}(x) - E(\hat{f}(x))] \\
&\quad + \sqrt{nh_1 \cdots h_q}\left[E(\hat{f}(x)) - f(x) - \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2 f_{ss}(x)\right] \\
&= \sqrt{nh_1 \cdots h_q}[\hat{f}(x) - E(\hat{f}(x))] \\
&\quad + O\left(\sqrt{nh_1 \cdots h_q}\sum_{s=1}^{q} h_s^3\right) \text{(根据式(1.36))} \\
&= \sum_{i=1}^{n} (nh_1 \cdots h_q)^{-1/2} \\
&\quad \times \left[K\left(\frac{X_i - x}{h}\right) - E\left(K\left(\frac{X_i - x}{h}\right)\right)\right] + o(1) \\
&\equiv \sum_{i=1}^{n} Z_{n,i} + o(1) \xrightarrow{d} N(0, \kappa^q f(x))
\end{aligned}$$

只要我们能证明Liapunov中心极限定理条件(A.21)成立,其中 30

$$Z_{n,i} = (nh_1 \cdots h_q)^{-1/2}\left[K\left(\frac{X_i - x}{h}\right) - E\left(K\left(\frac{X_i - x}{h}\right)\right)\right]$$

且根据式(1.37)

$$\sum_{i=1}^{n} \sigma_{n,i}^2 \stackrel{\text{def}}{=} \sum_{i=1}^{n} \mathrm{var}(Z_{n,i}) = \kappa^q f(x) + o(1)$$

Pagan 和 Ullah(1999, p. 40) 证明式(A. 21)在定理 1.3 给定的条件下成立。Pagan和 Ullah 使用的条件对某些 $\delta>0$ 有 $\int k(v)^{2+\delta}\mathrm{d}v<\infty$ 是被我们的假定 $k(v)$ 是非负和有界所隐含的,并且 $\int k(v)\mathrm{d}v=1$,因为 $\int k(v)^{2+\delta}\mathrm{d}v\leqslant C\int k(v)\mathrm{d}v=C$ 是有限的,其中 $C=\sup_{v\in\mathbb{R}^q}k(v)^{1+\delta}$。 □

1.10 一致收敛速度

到现在为止我们已经证明的只有逐点的和积分均方误差一致的(意味着依概率相合性)情形。在这一节我们扩展逐点收敛,以得到一个更强的"一致相合性"(uniform consistency)结果。我们将证明非参数核估计量是几乎确定一致收敛的(uniformly almost surely consistent),并且推导它们的几乎处处一致收敛速度。几乎处处一致收敛意味着依概率收敛;然而,反过来不成立,也就是,依概率收敛可能不意味着几乎处处收敛,参见 Serfling(1980)的具体例子。

我们已经对在 X 支撑的一个内点证实了逐点收敛。然而,当 x 在其支撑的边界时,诸如式(1.9)普通的核函数被证明可能不会产生 $f(x)$ 的一致估计,因此当考虑一致收敛速度时,我们需要排除边界范围。这强调了一般核估计的一个重要方面,并且在后面章节介绍的许多核估计量的动机就是消除这类"边界效应"(boundary effect)。我们首先证明当 x 是在(或接近)它支撑的边界时,$\hat{f}(x)$ 可能不是 $f(x)$ 的一个一致估计量。

考虑 X 是具有有界支撑的单变量情形。为了简便我们假定 $X\in[0,1]$。以前得到的逐点收敛结果 $\hat{f}(x)-f(x)=o_p(1)$ 要求 x 位于它的支撑内部。习题 1.13 证明,对位于其支撑边界的 x,$\mathrm{MSE}(\hat{f}(x))$ 可能不是 $o(1)$。因此,为了获得位于其支撑边界 x 的 $f(x)$ 的一致估计,可能需要作一些修正。典型的修正包括使用边界核(boundary kernel)和数据反射(data reflection)(见 Gasser 和 Muller(1979),Hall 和 Wehrly(1991),以及 Scott(1992, pp. 148—149))。作为例子,考虑 x 位于其最下边界的情况,即 $x=0$,因此,$\hat{f}(0)=(nh)^{-1}\sum_{i=1}^{n}K((X_i-0)/h)$。习题 1.13 证明在这种情况下,$E[\hat{f}(0)]=f(0)/2+O(h)$。因此,$\mathrm{bias}[\hat{f}(0)]=E[\hat{f}(0)]-f(0)=-f(0)/2+O(h)$,如果 $f(0)\neq 0$(当 $f(0)>0$ 时),这将不收敛于 0。

在文献中,各种边界核被提出以克服边界(偏误)问题。例如,一个简单的边界修正核由下式给出(假定 $X\in[0,1]$):

$$k_h(x,y) = \begin{cases} h^{-1}k\left(\dfrac{y-x}{h}\right)\Big/\int_{-x/h}^{\infty}k(v)\,\mathrm{d}v, & x \in [0,h) \\ h^{-1}k\left(\dfrac{y-x}{h}\right), & x \in [h,1-h] \\ h^{-1}k\left(\dfrac{y-x}{h}\right)\Big/\int_{-\infty}^{(1-x)/h}k(v)\,\mathrm{d}v, & x \in (1-h,1] \end{cases} \tag{1.43}$$

其中 $k(\cdot)$ 是满足式(1.10)的二阶核。现在,我们用下式估计 $f(x)$:

$$\hat{f}(x) = \frac{1}{n}\sum_{i=1}^{n}k_h(x,X_i) \tag{1.44}$$

其中 $k_h(x,X_i)$ 是定义在式(1.43)的。习题1.14证明上面的修正边界核成功地克服了边界问题。

我们现在确定对 $x \in \mathcal{S}$ 时 $\hat{f}(x)-f(x)$ 的几乎确定一致收敛速度,其中 $\mathcal{S}$ 是排除 X 支撑边界范围的有界集。在上面的例子中,当 x 的支撑是 $[0,1]$ 时,对任意小的正 $\varepsilon(0<\varepsilon<1/2)$ 我们可以选择 $\mathcal{S}=[\varepsilon,1-\varepsilon]$。我们假定 $f(x)$ 在 $\mathcal{S}$ 上是有下界的,大于某一个正常数。

定理 1.4 在 Masry(1996b)给出的关于 $f(\cdot)$ 的平滑条件下,并且假定 $\inf_{x\in\mathcal{S}}f(x) \geq \delta > 0$,我们几乎确定有

$$\sup_{x\in\mathcal{S}}|\hat{f}(x)-f(x)| = O\left(\frac{(\ln(n))^{1/2}}{(nh_1\cdots h_q)^{1/2}}+\sum_{s=1}^{q}h_s^2\right)$$

定理1.4的一个详细证明在1.12节给出。

32

因为几乎确定收敛意味着依概率收敛,收敛速度在概率上也成立,即在与定理1.4同样的条件下,我们有

$$\sup_{x\in\mathcal{S}}|\hat{f}(x)-f(x)| = O_p\left(\frac{(\ln(n))^{1/2}}{(nh_1\cdots h_q)^{1/2}}+\sum_{s=1}^{q}h_s^2\right)$$

使用式(1.36)和式(1.37)的结果,我们能确定如下的收敛速度。

定理 1.5 假定 $f(x)$ 为二阶导数有界的二次可微函数,那么我们有

$$\sup_{x\in\mathcal{S}}E\{[\hat{f}(x)-f(x)]^2\} = O\left(\sum_{s=1}^{q}h_s^4+(nh_1\cdots h_q)^{-1}\right)$$

证明:这从式(1.36)和式(1.37)得出,注意 $\sup_{x\in\mathcal{S}}f(x)$ 和 $\sup_{x\in\mathcal{S}}|f_{ss}(x)|$ $(s=1,\cdots,q)$ 都是有限的。 □

注意尽管依均方误差收敛意味着依概率收敛,人们不能从定理1.5推出在概率上的一致收敛速度。这是因为

$$E\{\sup_{x\in\mathcal{S}}[\hat{f}(x)-f(x)]^2\} \neq \sup_{x\in\mathcal{S}}E[\hat{f}(x)-f(x)]^2$$

$$P[\sup_{x\in\mathcal{S}}|\hat{f}(x)-f(x)|>\varepsilon]\neq\sup_{x\in\mathcal{S}}P[|\hat{f}(x)-f(x)|>\varepsilon]$$

这里的 sup 和 $E(\cdot)$ 或 $P(\cdot)$ 算子是不能彼此替换的。

Cheng(1997)提出了能够实现自动边界修正和具有一些典型最优性质的不同的(局部线性)密度估计量。Cheng 也提出了一种基于数据的窗宽选择量(根据插入法的思想),并且证明无论在密度支撑中是否有非平滑边界,这个窗宽选择量都是非常有效的。

33

1.11 高阶核函数

回忆一下减少 h 会降低一个核估计量的偏误但是增加它的方差。高阶核函数是用于减少偏误同时能够降低相应估计量均方误差的工具。许多流行的核函数(例如定义于式(1.10)中的)被称作"二阶"核。一个核的阶,$\nu(\nu>0)$,被定义为第一个非零矩的阶。例如,如果 $\int uk(u)\mathrm{d}u=0$,但是 $\int u^2k(u)\mathrm{d}u\neq0$,那么 $k(\cdot)$ 被称作一个二阶核($\nu=2$)。因此一个一般的 ν 阶核($\nu\geq2$ 是一个整数)必须满足如下条件:

$$\begin{aligned}&(\text{i})\ \int k(u)\mathrm{d}u=1\\&(\text{ii})\ \int u^lk(u)\mathrm{d}u=0,\quad l=1,\cdots,\nu-1\\&(\text{iii})\ \int u^\nu k(u)\mathrm{d}u=\kappa_\nu\neq0\end{aligned}\tag{1.45}$$

显然,当 $\nu=2$ 时,式(1.45)退化为式(1.10)。

如果人们用一个 ν 阶核函数取代式(1.35)中的 $\hat{f}(x)$ 的二阶核,那么,与使用一个二阶核的情况一样,在 $f(x)$ 是 ν 阶可微的假定下,以及假定所有 h_s 有相同大小的阶,可以证明

$$\operatorname{bias}(\hat{f}(x))=O\left(\sum_{s=1}^q h_s^\nu\right)\tag{1.46}$$

$$\operatorname{var}(\hat{f}(x))=O((nh_1\cdots h_q)^{-1})\tag{1.47}$$

(见习题 1.15)因此,我们有

$$\operatorname{MSE}(\hat{f}(x))=O\left(\sum_{s=1}^q h_s^{2\nu}+(nh_1\cdots h_q)^{-1}\right)\tag{1.48}$$

$$\hat{f}(x)-f(x)=O_p\left(\sum_{s=1}^q h_s^\nu+(nh_1\cdots h_q)^{-1/2}\right)$$

因此,通过使用一个 ν 次高阶核函数($\nu>2$),能够把 $\hat{f}(x)$ 的偏误的阶从

$O\left(\sum_{s=1}^{q} h_s^2\right)$减少到$O\left(\sum_{s=1}^{q} h_s^{\nu}\right)$，并且$h_s$的最优值可通过平衡平方偏误和方差得到，给定$h_s = O(n^{-1/(2\nu+q)})$，其中收敛速度现在变为$\hat{f}(x) - f(x) = O_p(n^{-\nu/(2\nu+q)})$。假定$f(x)$在任何有限阶都是可微的，那么人们能够选择足够大的$\nu$，并且相应的速度可以任意接近于$O_p(n^{-1/2})$。然而，注意对$\nu > 2$，没有满足式(1.45)的非负核存在。这必然意味着，我们必须指定负的权重给某个范围的数据，这隐含着人们可能得到负的密度估计，这是一个明显不被期望的负效应。进一步，在有限样本应用中非负二阶核经常被发现能比它们相应更高阶的核产生更稳定的估计结果。因此，高阶核函数主要被用于理论的目的；例如，在一个半参数模型中为了得到一些有限维参数的$\sqrt{n}$收敛速度，人们经常必须使用高阶核函数（这样的一个例子见第7章）。 34

高阶核函数是很容易构造的。假定$k(u)$围绕零对称[①]，即$k(u) = k(-u)$，那么对所有正整数m有$\int u^{2m+1}k(u)\mathrm{d}u = 0$。作为例子，为构建一个简单的四阶核（即$\nu = 4$），人们可以从一个二阶核例如标准正态核开始，在它的变量中设定一个多项式，并求解这个受到矩约束的多项式的根。例如，令$\Phi(u) = (2\pi)^{-1/2}\exp(-u^2/2)$是一个二阶高斯核，我们可以从如下多项式开始：

$$k(u) = (a + bu^2)\Phi(u) \tag{1.49}$$

其中a和b是必须满足一个四阶核条件的两个常数。令$k(u)$满足式(1.45)且有$\nu = 4\left(\int u^l k(u)\mathrm{d}u = 0, l = 1,3,\text{因为}k(u)\text{是一个偶函数}\right)$，我们因此仅仅要求$\int k(u)\mathrm{d}u = 1$和$\int u^2 k(u)\mathrm{d}u = 0$。从这两个约束条件，能够很容易得到结果$a = 3/2$和$b = -1/2$。对于需要更高阶核函数的读者，我们提供了一些基于二阶高斯和Epanechnikov核的例子，它们可能是在应用非参数估计中两个最流行的核。正如四阶单变量高斯核由以下公式给出： 35

$$k(u) = \left(\frac{3}{2} - \frac{1}{2}u^2\right)\frac{\exp(-u^2/2)}{\sqrt{2\pi}}$$

而六阶单变量高斯核由下式给出：

$$k(u) = \left(\frac{15}{8} - \frac{5}{4}u^2 + \frac{1}{8}u^4\right)\frac{\exp(-u^2/2)}{\sqrt{2\pi}}$$

二阶单变量Epanechnikov核是基于对最小化核估计量的积分均方误差的一个变分解的最优核（见Serfling(1980, pp. 40—43)）。单变量二阶Epanechnikov核

---

① 一般只有对称核函数被用于实践中，但可参见Abadir和Lawford(2004)关于最优非对称核的近期工作。

由下式给出：

$$k(u)=\begin{cases}\dfrac{3}{4\sqrt{5}}\left(1-\dfrac{1}{5}u^2\right), & u^2<5.0\\ 0, & \text{其他}\end{cases}$$

四阶单变量 Epanechnikov 核由下式给出：

$$k(u)=\begin{cases}\dfrac{3}{4\sqrt{5}}\left(\dfrac{15}{8}-\dfrac{7}{8}u^2\right)\left(1-\dfrac{1}{5}u^2\right), & u^2<5.0\\ 0, & \text{其他}\end{cases}$$

其中六阶单变量 Epanechnikov 核由下式给出：

$$k(u)=\begin{cases}\dfrac{3}{4\sqrt{5}}\left(\dfrac{175}{64}-\dfrac{105}{32}u^2+\dfrac{231}{320}u^4\right)\left(1-\dfrac{1}{5}u^2\right), & u^2<5.0\\ 0, & \text{其他}\end{cases}$$

图 1.2 画出了上面定义的二阶、四阶和六阶 Epanechnikov 核。显然，对 $\nu>2$，这些核的确指定了可能引起负密度估计的负权重，这不是一个期望的特征。

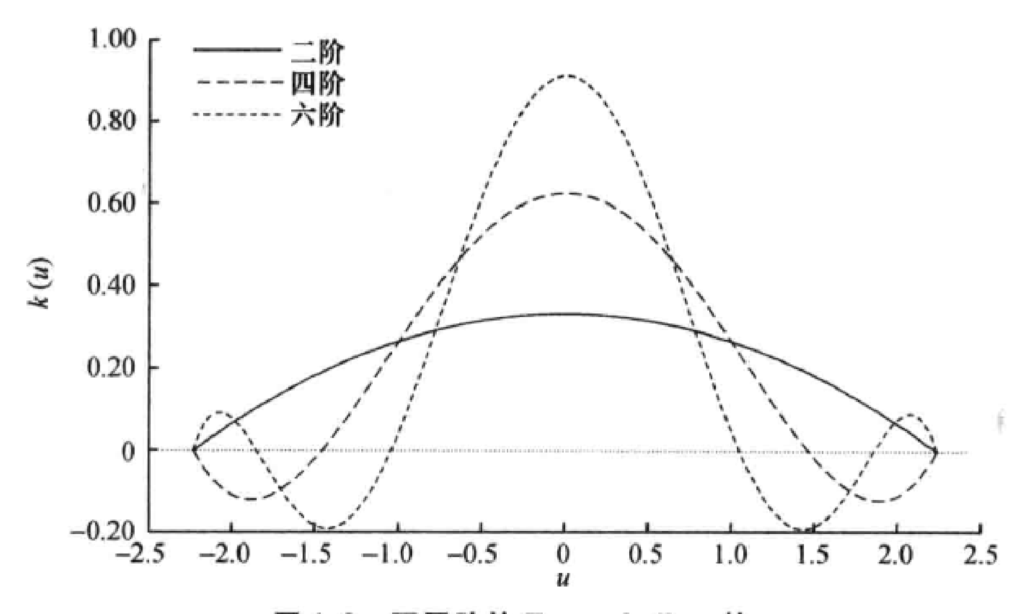


图 1.2　不同阶的 Epanechnikov 核

至于在单变量核密度估计背景下对高阶核准确的积分均方误差的相关研究，参见 Hansen(2005)。此外，对使用迭代法来估计转换核密度的相关研究，参见 Yang 和 Marron(1999)以及 Yang(2000)。

1.12　定理 1.4 的证明

下面的证明是基于 Masry(1996b)中给出的论证，他建立了对带有弱依赖

(α 混合)数据的局部多项式回归的几乎确定一致收敛速度;关于弱依赖过程的进一步细节见第18章。由于核密度估计量的偏误是 $O\left(\sum_{s=1}^{q} h_s^2\right)$ 阶并且方差是 $O((nh_1\cdots h_q)^{-1})$ 阶,很容易证明最优收敛速度要求所有的 h_s 应该有相同大小的阶。因此,为了记号的简洁,我们使用简化的假定:

36

$$h_1 = \cdots = h_q = h$$

这不会影响最优收敛速度,但是它极大地简化了推导。我们强调,在应用中,人们应该总是允许 $h_s(s=1,\cdots,s)$ 彼此不同,当使用例如交错鉴定等完全数据驱动窗宽选择方法时,这当然是一直被允许的。我们仅仅对紧接着的理论分析假定全部平滑参数是相同的。

证明:令 $W_n = W_n(x) = |\hat{f}(x) - f(x)|$。为证明随机变量 W_n 几乎确定是 η 阶的,我们能够证明 $\sum_{n=1}^{\infty} P(|W_n/\eta| > 1)$ 是有限的(某些 $\eta>0$)。那么,根据 Borel-Cantelli 引理(见附录 A 引理 A.7),我们知道几乎确定 $W_n = O(\eta)$。这里,上确界算子使得证明复杂化,因为$\mathcal{S}$是一个不可数集。令 L_n 表示一个可数集,那么,我们有

37

$$P(\max_{x\in L_n} W_n(x) > \eta) \leqslant (\# L_n) \max_{x\in L_n} P(W_n(x) > \eta) \tag{1.50}$$

但在我们的情形中,$x\in\mathcal{S}$是不可数的,并且为了给出 $P(\sup_{x\in\mathcal{S}} W_n(x) > \eta)$ 的上界,我们不能简单地使用类似式(1.50)的不等式。

然而,由于$\mathcal{S}$是有界集,我们能够把$\mathcal{S}$分割为许多可数的子集,使每个子集规模尽可能小。那么 $P(\sup_{x\in\mathcal{S}} |W_n(x)| > \eta)$ 能够被转换为类似 $P(\max_{x\in L_n} |W_n(x)| > \eta)$ 的一个问题,并且不等式(1.50)能够被用于处理这一项。我们使用这种思路证明定理1.4。

我们记

$$\begin{aligned} |\hat{f}(x) - f(x)| &= |\hat{f}(x) - E(\hat{f}(x)) + E(\hat{f}(x)) - f(x)| \\ &\leqslant |\hat{f}(x) - E(\hat{f}(x))| + |E(\hat{f}(x)) - f(x)| \end{aligned}$$

我们通过证明下式证明定理1.4:

$$\sup_{x\in\mathcal{S}} |E(\hat{f}(x)) - f(x)| = O(h^2) \tag{1.51}$$

且

$$\sup_{x\in\mathcal{S}} |\hat{f}(x) - E(\hat{f}(x))| = O\left(\frac{(\ln(n))^{1/2}}{(nh^q)^{1/2}}\right) \text{ a.s.} \tag{1.52}$$

我们首先证明式(1.51)。由于紧集$\mathcal{S}$是在其支撑的内部,通过一个变量替换后的变量(change-of-variables),我们有,对 $x\in\mathcal{S}$,

$$E(\hat{f}(x)) - f(x) = \int f(x+hv)K(v)\mathrm{d}v - f(x)$$
$$= h^2\int v'f^{(2)}(\tilde{x})vK(v)\mathrm{d}v$$
$$\leqslant C_1h^2\int v'vK(v)\mathrm{d}v \leqslant Ch^2 = O(h^2)$$

在 $x\in\mathcal{S}$一致。因此,我们证明了式(1.51)。

现在我们转向式(1.52)的证明。由于$\mathcal{S}$是紧致的(闭且有界),它可被以 38 $x_{k,n}$为中心且长度为 $l_n(k=1,\cdots,L(n))$的(q 维)立方 $I_k=I_{k,n}$中的一个有限数 $L_n=L(n)$覆盖。因为$\mathcal{S}$是紧致的,我们知道 $L_n=$常数$/(l_n)^q$,这给出 $l_n=$常数$/L_n^{1/q}$。我们记

$$\sup_{x\in\mathcal{S}}|\hat{f}(x)-E(\hat{f}(x))| = \max_{1\leqslant k\leqslant L(n)}\sup_{x\in\mathcal{S}\cap I_k}|\hat{f}(x)-E(\hat{f}(x))|$$
$$\leqslant \max_{1\leqslant k\leqslant L(n)}\sup_{x\in\mathcal{S}\cap I_k}|\hat{f}(x)-\hat{f}(x_{k,n})|$$
$$+\max_{1\leqslant k\leqslant L(n)}|\hat{f}(x_{k,n})-E(\hat{f}(x_{k,n}))|$$
$$+\max_{1\leqslant k\leqslant L(n)}\sup_{x\in\mathcal{S}\cap I_k}|E(\hat{f}(x_{k,n}))-E(\hat{f}(x))|$$
$$\equiv Q_1+Q_2+Q_3$$

注意 Q_2 不依赖于 x,因此,$\sup_{x\in\mathcal{S}\cap I_k}$ 没有出现在 Q_2 的定义中。

我们首先考虑 Q_2。记 $W_n(x) = \hat{f}(x) - E(\hat{f}(x)) = \sum_i Z_{n,i}$,其中 $Z_{n,i} = (nh^q)^{-1}\{K((X_i-x)/h) - E[K((X_i-x)/h)]\}$。对任意 $\eta>0$,我们有

$$P[Q_2>\eta] = P[\max_{1\leqslant k\leqslant L(n)}|W_n(x_{k,n})|>\eta]$$
$$\leqslant P[W_n(x_{1,n})>\eta \text{ 或 } W_n(x_{2,n})>\eta,\cdots,\text{或 } W_n(x_{L(n),n})>\eta]$$
$$\leqslant P(W_n(x_{1,n})>\eta)+P(W_n(x_{2,n})>\eta)+\cdots$$
$$+P(W_n(x_{L(n),n})>\eta)$$
$$\leqslant L(n)\sup_{x\in\mathcal{S}}P[|W_n(x)|>\eta] \tag{1.53}$$

因为 $K(\cdot)$是有界的,并令 $A_1=\sup_x|K(x)|$,我们有 $|Z_{n,i}|\leqslant 2A_1/(nh^q)$,所有 $i=1,\cdots,n$。定义 $\lambda_n=(nh^q\ln(n))^{1/2}$。那么当 n 足够大时,对所有 $i=1,\cdots,n$ 有 $\lambda_n|Z_{n,i}|\leqslant 2A_1[\ln(n)/(nh^q)]^{1/2}\leqslant 1/2$。[①] 使用不等式 $\exp(x)\leqslant 1+x+x^2$,

① 现在,任何 $\lambda_n\leqslant(nh^q)/(4A_1)$将导致$|\lambda_nZ_{n,i}|\leqslant 1/2$。稍后我们将证明,为得到 Q_2 的最优率,人们需要选择 $\lambda_n=(nh^q\ln(n))^{1/2}$。

$|x|\leqslant 1/2$,我们有 $\exp(\pm\lambda_n Z_{n,i})\leqslant 1\pm\lambda_n Z_{n,i}+\lambda_n^2 Z_{n,i}^2$。因此,

$$E[\exp(\pm\lambda_n Z_{n,i})]\leqslant 1+\lambda_n^2 E[Z_{n,i}^2]\leqslant \exp[E(\lambda_n^2 Z_{n,i}^2)] \tag{1.54}$$

其中我们使用 $E(Z_{n,i})=0$,而对第二个不等式我们使用 $1+v\leqslant \exp(v)$,$v\geqslant 0$($v=E[\lambda_n^2 Z_{n,i}^2]$)。

根据 Markov 不等式(见引理 A.23 有 $\phi(x)=\exp(ax)$),我们知道

$$P[X>c]\leqslant \frac{E[\exp(Xa)]}{\exp(ac)},\quad a>0 \tag{1.55}$$

使用式(1.55)我们得到

$$\begin{aligned}
P[|W_n(x)|>\eta] &= P\left[\left|\sum_{i=1}^n Z_{n,i}\right|>\eta\right]\\
&= P\left[\sum_{i=1}^n Z_{n,i}>\eta\right]+P\left[\sum_{i=1}^n Z_{n,i}<-\eta\right]\\
&\leqslant P\left[\sum_{i=1}^n Z_{n,i}>\eta\right]+P\left[-\sum_{i=1}^n Z_{n,i}>\eta\right]\\
&\leqslant \frac{E\left[\exp\left(\lambda_n\sum_{i=1}^n Z_{n,i}\right)\right]+E\left[\exp\left(-\lambda_n\sum_{i=1}^n Z_{n,i}\right)\right]}{\exp(\lambda_n\eta_n)}\\
&\quad (\text{由式(1.55)}, a=\lambda_n, c=\eta)\\
&\leqslant 2\exp(-\lambda_n\eta)\left[\exp\left(\lambda_n^2\sum_{i=1}^n E(Z_{n,i}^2)\right)\right]\\
&\quad (\text{根据式(1.54)})\\
&\leqslant 2\exp(-\lambda_n\eta)[\exp(A_2\lambda_n^2/(nh^q))]
\end{aligned} \tag{1.56}$$

其中我们使用

$$E[Z_{n,i}^2]\leqslant (nh^q)^{-2}E[K^2((X_i-x)/h)]\leqslant A_2(n^2h^q)^{-1}[1+o(1)]$$

因为式(1.56)中最后的上界独立于 x,它也是一致有界,即

$$\sup_{x\in\mathcal{S}} P[|W_n(x)|>\eta]\leqslant 2\exp\left(-\lambda_n\eta+\frac{A_2\lambda_n^2}{nh^q}\right) \tag{1.57}$$

我们希望 $\eta\to 0$ 尽可能快,并且同时我们需要 $\lambda_n\eta\to\infty$ 的速度能保证式(1.57)可求和。[①] 我们能够选择 $\lambda_n\eta=C_4\ln(n)$,或 $\lambda_n=C_4\ln(n)/\eta$。找出 $\eta\to 0$ 的最快速度等价于找出 $\lambda_n\to\infty$ 的最快速度。我们也需要 $\lambda_n\eta$ 的阶 $\geqslant\lambda_n^2/$

---

① 如果 $\left|\sum_{j=1}^{\infty}a_j\right|<\infty$,序列 $\{a_n\}_{n=1}^{\infty}$ 被称作可求和的。

40 (nh^q),或 $\ln(n) \geqslant \lambda_n^2/(nh^q)$。因此,我们仅仅需要在约束 $\lambda_n^2 \leqslant (nh^q)\ln(n)$ 下最大化 $\lambda_n \to \infty$ 的阶。这样做,我们得到

$$\lambda_n = [(nh^q)\ln(n)]^{1/2}, \eta = C_4\ln(n)/\lambda_n = C_4[\ln(n)/(nh^q)]^{1/2} \tag{1.58}$$

使用式(1.58),我们有

$$-\lambda_n\eta + A_2\lambda_n^2/(nh^q) = -C_4\ln(n) + A_2\ln(n) = -\alpha\ln(n)$$

其中 $\alpha = C_4 - A_2$。把该式代入式(1.57),然后代入式(1.53),我们有

$$P[Q_2 > \eta_n] \leqslant 2L(n)/n^\alpha \tag{1.59}$$

通过选择足够大的 C_4,我们再选择恰当的 $L(n)$ 的阶,能够得到 $L(n)/n^\alpha$ 是可求和的结果,即 $\sum_{n=1}^{\infty} P(|Q_2/\eta_n| > 1) \leqslant 4\sum_{n=1}^{\infty} L(n)/n^\alpha < \infty$。因此,根据 Borel-Cantelli 引理,我们知道

$$Q_2 = O(\eta_n) = O(\ln(n)^{1/2}/(nh^q)^{1/2}) \quad \text{a.s.} \tag{1.60}$$

我们现在考虑 Q_1 和 Q_3。回忆 $\|\cdot\|$ 表示一个向量的普通欧几里得范数。根据关于 $K(\cdot)$ 的 Lipschitz 条件,我们知道

$$\sup_{x\in S\cap I_k} |K((X_i - x)/h) - K((X_i - x_{k,n})/h)| \leqslant C_1h^{-1}\sup_{x\in S\cap I_k}\|x - x_{k,n}\| \leqslant C_2h^{-1}l_n$$

因此,通过选择 $l_n = (\ln(n))^{1/2}h^{(q+2)/2}/n^{1/2}$,我们有

$$|Q_1| \leqslant C_2h^{-(q+1)}l_n = O((\ln(n)/(nh^q))^{1/2}) \tag{1.61}$$

通过完全相同的论证我们能够证明

$$|Q_3| \leqslant C_3h^{-(q+1)}l_n = O((\ln(n)/(nh^q))^{1/2}) \tag{1.62}$$

方程(1.60)到(1.62)证明了式(1.52),并且这完成了定理 1.4 的证明。

□

1.13 应用

我们现在考虑一些能够说明核方法灵活性和作用的单变量和多变量密度估计的应用。

41

1.13.1 女性工资不平等

DiNardo 和 Tobias(2001,p.12)使用非参数核方法来研究从 1979 年到 1989 年的女性工资不平等现象。有时一个参数分布的尺度被用作不平等的一个

原始测度，并且对数工资的标准差在这段时期从 0.41 增加到 0.50，提高了 25%。[①] 人们可能认为这些变化背后共同的原因将包括国际贸易、技术变化或者可能的组织变迁。正如我们将在下面看到的，DiNardo 和 Tobias 证明核估计量有助于说明什么是真正的原因。

如果使用一个参数模型并且假定，比如说，对数工资服从一个正态分布，人们将获得出现在图 1.3 中的数据描述。

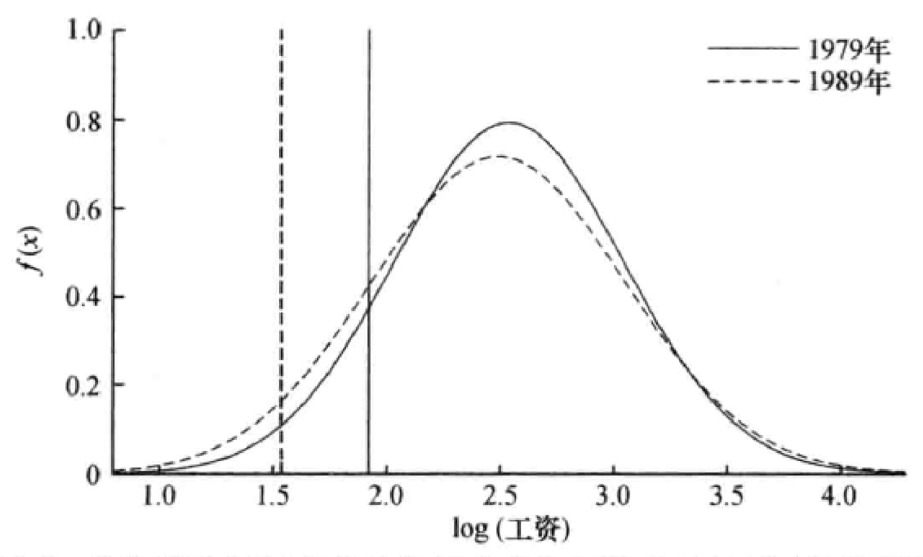


图 1.3 参数密度估计（竖线代表 1979 年和 1989 年的（对数）最低工资）

使用非参数核方法和一个采用二阶高斯核的简单“正态推断经验法则”($h = 1.06\sigma n^{-1/5}$)窗宽产生了图 1.4 中画出的估计量。 42

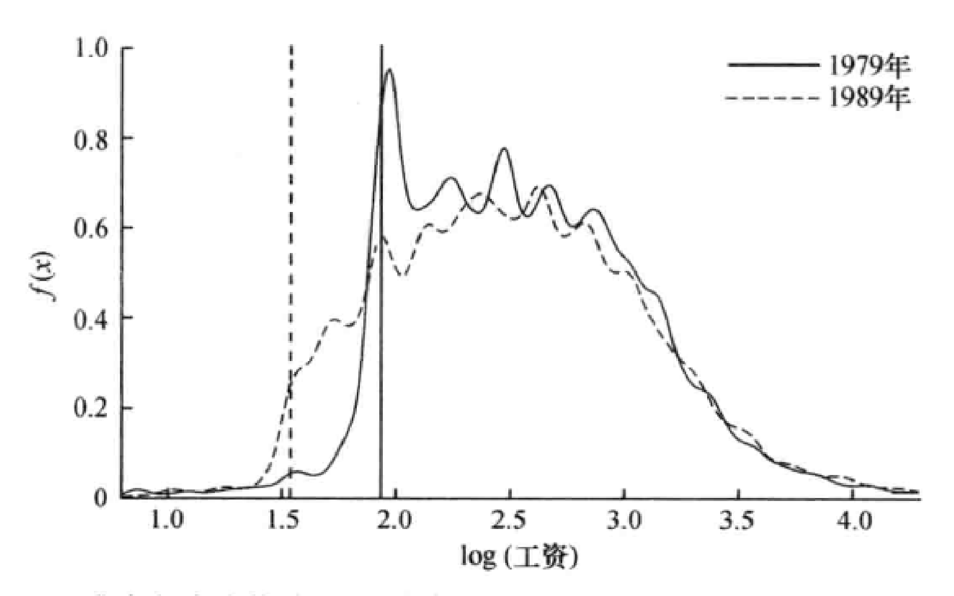


图 1.4 非参数密度估计（竖线代表 1979 年和 1989 年的（对数）最低工资）

① 在 1979 年和 1989 年的最低工资是 2.90 美元/小时和 3.35 美元/小时，而 CPI 在 1979 年、1989 年和 2000 年分别是 72.6、124.0 和 172.2。工资来自当前人口调查（Current Population Survey, CPS）。在 1979 年和 1989 年的样本中分别有 140 284 和 167 863 个观测值。高斯核被使用，并且正态推断经验法则窗宽对 1979 年和 1989 年的样本分别是 0.050 和 0.053。出现在图 1.3 和 1.4 中的工资值是当前（2000 年）美元价值。

出现在图 1.4 中的基于正态推断经验法则的两个核密度估计看起来是平滑不足的。然而,这些估计清楚地揭示了一个没有被参数方法捕捉的特征:一个在 1979 年有约束力的最低工资标准在 1989 年对大多数女性不再起约束作用。这个发现意味着工资不平等的增长能够被在 1979 年一个有约束力的真实最低工资所引起的截断来解释。也就是说,在 1979 年,不像在 1989 年,雇主对许多雇员支付最低工资,这扭曲并减少了工资分布的方差。最低工资真实值随着时间推移而下降,在 1989 年成为无约束力的。因此,非参数估计量已经很好地揭示了扩大的工资不平等背后的真实原因,并把注意力从其他可能的解释上移开,比如国际贸易、技术变化或者可能的组织变迁。这个例子只是用来强调传统参数方法可能掩盖数据中出现的重要特征这个事实。

43

1.13.2 失业率和城市规模

在这个例子中我们基于一个 $n=295$ 个城市的样本使用了关于城市人口(ln(城市规模))和失业率的美国数据。Gan 和 Zhang(2006)提出一个理论,预言城市规模越大,失业率越低(平均意义上)。在图 1.5 中,我们使用最小二乘交错窗宽选择和一个二阶高斯核画出了估计的联合概率密度函数。失业率和人口的交错鉴定窗宽分别是 0.665 和 0.351。

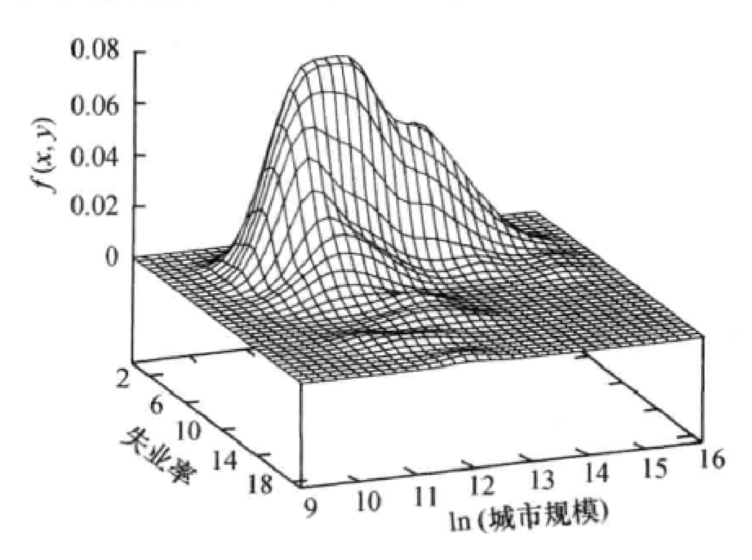


图 1.5 失业率和 ln(城市规模)的联合密度估计

出现在图 1.5 中的联合密度估计是跟大城市倾向于有低失业率并且反之亦然的假说一致的。也就是说,图 1.5 揭示了某种程度上的"直角分布",在低失业率和大的城市规模处有概率密度(probability mass),而当城市规模下降时我们观察到概率密度首先转移向原点,然后当城市规模进一步下降时,这个密度转移向更高的失业率。

1.13.3 青少年成长

青少年不正常的生长能够提供一个儿童有医疗问题的早期警告。例如,太快的生长可能暗示着脑水肿(在头颅内的积液)、脑瘤或其他会引起巨头症(有一个不正常的大脑袋)的条件的出现,而太慢的生长可能意味着大脑畸形、颅缝早期融合或其他问题。身高、体重增加不足或两者综合不足可能意味着生长迟滞、慢性疾病、未好好照管或其他问题。

我们研究来自美国疾病控制和预防中心(CDC)国家健康和营养检测调查关于美国孩子健康的人口数据。我们整合数据并使用1999/2000和2001/2002这两个有全国代表性的健康检测调查横截面数据。对每个横截面,两个单独的数据集必须被联系起来(一个身体检查数据集和一个人口变量数据集)。整合起来的联合数据集包括8 399个位于2—20岁年龄段的儿童和青少年的完整观测值。我们根据性别为身高和体重构建联合分布。

图1.6和图1.7显示男性和女性的身高和体重联合分布是类似的;然而,男性在更高的身高和体重值处有更大的概率密度。也就是说,人们更容易观测到又高又重的男孩而不是女孩。这类数据奠定了构造青少年成长图形的基础,比如身高体重图。[①] 也可参见Wei和He(2006)对有条件的生长图的相关研究。

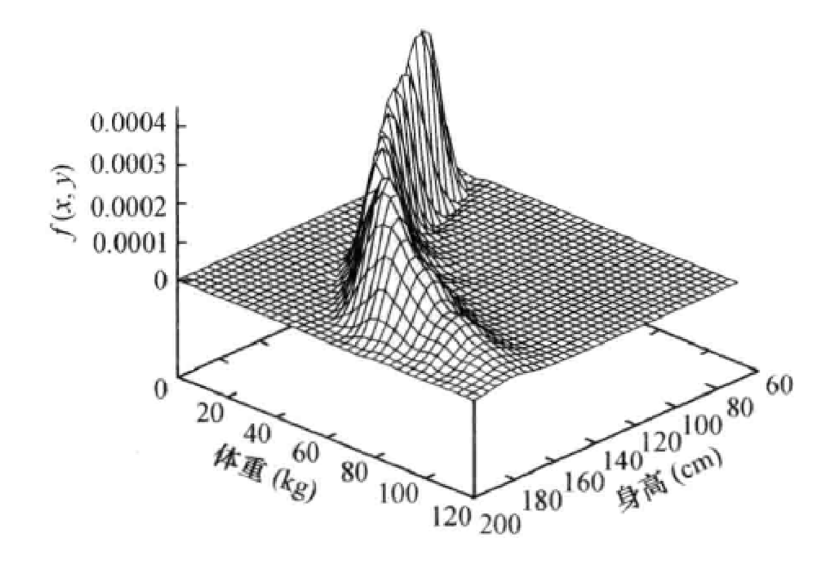


图1.6 男性体重和身高的联合密度估计

① 由美国国家健康统计中心制作的官方成长图见 http://www.cdc.gov/growthcharts。

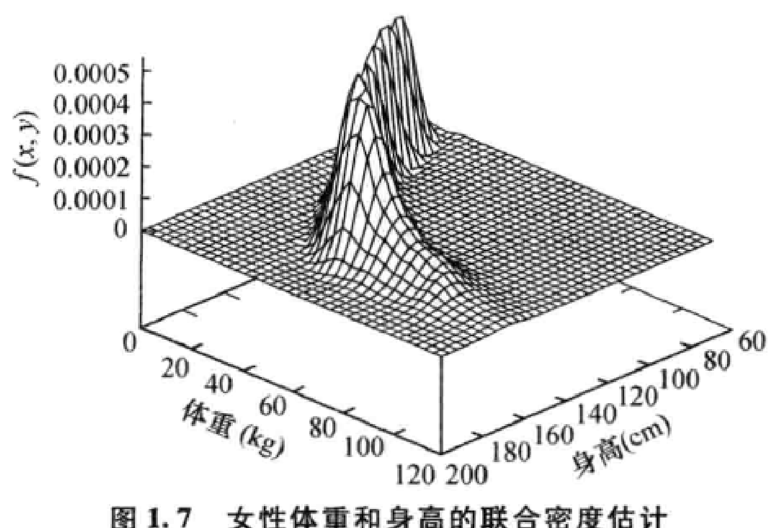


图 1.7 女性体重和身高的联合密度估计

1.13.4 老忠实喷泉数据

老忠实喷泉是位于黄石国家公园的一个旅游名胜。这个著名的数据集包括由 2 个变量构成的 $n=272$ 个观测值，喷发持续时间（分钟）和等到下一次喷发的时间（分钟）。这个数据集被公园服务部门用于建模，除了其他因素外，期望持续时间条件于自上次喷发后已经过去的时间量。然而，联合分布建模所关心的是自身的合理性。联合概率密度函数的潜在双峰特征由图 1.8 中给出的核估计量很好地揭示，它的构造使用了似然交错鉴定窗宽和一个二阶高斯核。①

45

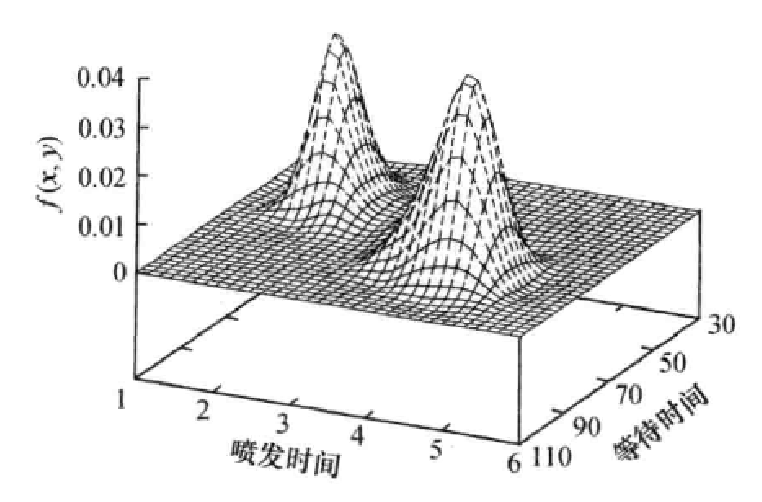


图 1.8 老忠实喷泉数据的联合密度估计

如果要用一个参数模型，例如二元正态（满足对称、单峰和远离峰时单调递减），来为这个密度建模，人们当然不能揭示被核估计很好显示的那种潜在结构。

① 似然交错鉴定窗宽计算等于$(h_1,h_2)=(0.368\sigma_1 n^{-1/6},0.764\sigma_2 n^{-1/6})$，而最小二乘交错鉴定窗宽是$(h_1,h_2)=(0.307\sigma_1 n^{-1/6},0.733\sigma_2 n^{-1/6})$，其中 h_1 是喷发持续时间的窗宽，而 h_2 是等待时间的窗宽。

1.13.5　意大利真实收入分布的演化，1951—1998 年

Baiocchi(2006)使用核方法研究了意大利真实收入分布的演化。他研究了 21 个地区收入分布的一系列“堆积的”(stacked)单变量核密度估计并画出了这些单变量核密度估计随着时间推移的相应演化。我们感谢 Giovanni Baiocchi 提供了使用似然交错鉴定用于产生一系列单变量核估计的包括 1951—1998 年(百万里拉，1990 = 基年)观测值的数据。图 1.9 通过在一个三维图中堆积年度(即横截面)单变量核估计呈现了真实人均 GDP 的演化(百万 1990 年里拉)。

46

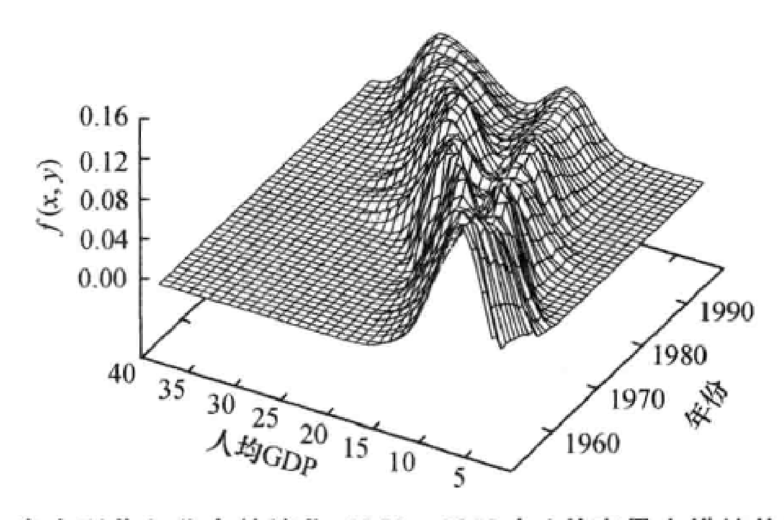


图 1.9　意大利收入分布的演化，1951—1998 年(单变量交错核估计序列)

图 1.9 显示收入分布已经从 20 世纪 50 年代初期的单峰分布演化为 20 世纪 90 年代明显的双峰分布。这个结果对窗宽选择是稳健的，并且不管用简单经验法则还是诸如似然交错鉴定之类的数据驱动方法都能被观测到。核方法很好地显示了这种演化，如果人们使用收入分布的参数模型，这种演化可能很容易被忽略(比如在应用研究中广泛使用的对数正态分布)。

47

1.14　习　题

习题 1.1　考虑如下连续数据样本：

$\{-0.57, 0.25, -0.08, 1.40, -1.05, -1.00, 0.37, -1.15, 0.73, 1.59\}$

(例如，以万亿美元计的真实 GDP 季节调整缺口)。

回忆由下式给出的参数正态概率密度函数：

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

(i) 假定一个潜在正态分布，计算这些数据并画出参数密度函数(即计算 $\hat{\mu}$ 和 $\hat{\sigma}^2$)。

48 (ii) 从 -1.5 到 2.0 使用 0.5 的箱宽计算这些数据并画出直方图。

回忆连续数据的一个单变量密度函数的核估计量可以被表示为

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)$$

并且一个常用(最优)核是 Epanechnikov 核,由下式给出:

$$K\left(\frac{X_i - x}{h}\right) = \begin{cases} \frac{3}{4\sqrt{5}}\left(1 - \frac{1}{5}\left(\frac{X_i - x}{h}\right)^2\right), & \left|\frac{X_i - x}{h}\right| < \sqrt{5} \\ 0, & \text{其他} \end{cases}$$

其中 h 是一个平滑参数,受到位于范围$(0,\infty]$的限制。

49 (iii) 使用同样小的数据样本,使用窗宽 $h=1.5$,对每个样本实现计算密度函数的核估计量。写出全部步骤。

(iv) 使用同样的数据,使用窗宽 $h=1.5$,对每个样本实现计算密度函数的核估计量。写出全部步骤。

(v) 在同样的轴上,对每个函数使用一个平滑曲线来"连接点"(connect the dots),画出你的密度函数估计量。

(vi) 描述增加平滑参数对估计的密度函数的影响。

习题 1.2 令 $\hat{p}$ 是定义于式(1.1)的。证明 $\hat{p}$ 是 $p=P(H)$ 的极大似然估计量。

提示:如果第 i 次试验是 H,定义 $X_i=1$,如果是 T,则定义 $X_i=0$。那么似然函数为 $\prod_{i=1}^{n} f(X_i) = \prod_{i=1}^{n} p^{X_i}(1-p)^{1-X_i}$。对数似然函数为 $L = \left(\sum_{i=1}^{n} X_i\right)\ln p + \left(\sum_{i=1}^{n}(1-X_i)\right)\ln(1-p)$。

习题 1.3

(i) 证明 $\mathrm{MSE}(\hat{p}_n) = p(1-p)/n$,其中 $p=P(H)$。

(ii) 证明$\mathrm{plim}_{n\to\infty}\hat{p}_n = p$。

(iii) 假定 $p=P(H)\in(0,1)$,证明通常极限$\lim\limits_{n\to\infty}\hat{p}_n$ 不存在。

注意极限被定义如下:令 a_n 为一实数序列,我们记$\lim\limits_{n\to\infty} a_n = c$,如果对所有任意小 $\varepsilon>0$,存在一个正整数 n_0,满足:对所有 $n\geqslant n_0$, $|a_n - c| < \varepsilon$。

提示:对(ii)使用(i)的结果和附录 A 的定理 A.3。

提示:对(iii)通过反证法证明。

习题 1.4 令 $F_n(x)$是定义在式(1.2)中的。

(i) 证明 $\mathrm{MSE}\{F_n(x)\} = O(n^{-1})$(注意根据附录 A 的定理 A.7,这意味着 $F_n(x) - F(x) = O_p(n^{-1/2})$)。

(ii) 证明

50

$$\sqrt{n}(F_n(x)-F(x)) \xrightarrow{d} N(0,F(x)(1-F(x)))$$

提示：首先证明 $E[F_n(x)]=F(x)$ 以及 $\mathrm{var}(F_n(x))=n^{-1}F(x)(1-F(x))$，然后使用 Lindeberg-Levy 中心极限定理。

习题 1.5　在假定 $f(x)$ 在 x 处有连续二阶导数下证明式(1.13)。

提示：使用由附录 A 引理 A.13 给出的占优收敛定理。

习题 1.6　记 $k_{ij}=k\left(\frac{X_i-X_j}{h}\right)$ 和 $\bar{k}_{ij}=\bar{k}\left(\frac{X_i-X_j}{h}\right)$。使用 $n^{-2}=(n(n-1))^{-1}-(n^2(n-1))^{-1}$，我们由式(1.23)得到

$$\begin{aligned}
\mathrm{CV}_f(h) &= \frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{k}_{ij}-\frac{2}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}^{n}k_{ij} \\
&\quad -\frac{1}{n^2(n-1)h}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{k}_{ij} \\
&= \frac{1}{(n-1)h}\bar{k}(0)+\frac{1}{n(n-1)h}\sum_{i=1}^{n}\sum_{j\neq i}^{n}[\bar{k}_{ij}-2k_{ij}]+O_p(n^{-1}) \\
&= \frac{\kappa}{(n-1)h}+J_n+O_p(h(nh)^{-1}) \qquad (1.63)
\end{aligned}$$

其中 $J_n=[n(n-1)h]^{-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n}[\bar{k}_{ij}-2k_{ij}]$ 且 $\kappa=\int k^2(v)\mathrm{d}v\equiv\bar{k}(0)$。

(i) 证明 $E(J_n)=B_0+B_1h^4+O(h^5)$，其中 $B_0=-\int f(x)^2\mathrm{d}x$ 且 $B_1=(\kappa_2^2/4)\left\{\int[f^{(2)}(x)]^2\mathrm{d}x\right\}$。

(ii) 接受 $J_n=E(J_n)$ + 低阶项这个事实，因此，最小化 $\mathrm{CV}_f(h)$ 渐近地等价于最小化 $I(h)\overset{\mathrm{def}}{=}(nh)^{-1}\kappa+E(J_n)$。得到最小化 $I(h)$ 的 $\hat{h}$。

(iii) 假定对所有 v 有 $k(0)\geqslant k(v)$（这对核估计通常是成立的）。如果我们不用去一估计量，那么我们将有目标函数 $V(h)\overset{\mathrm{def}}{=}(nh)^{-1}[\kappa-2k(0)]+E(J_n)$。证明 $h=0$ 最小化 $V(h)$，这显然违反了条件当 $n\to\infty$ 时，$nh\to\infty$。这说明当构造 $\mathrm{CV}_f(h)$ 时我们必须使用去一估计量。

(iv) 在推导式(1.63)中我们使用了

51

$$A_n\overset{\mathrm{def}}{=}[n^2(n-1)h]^{-1}\sum_{i}\sum_{j\neq i}\bar{k}((X_i-X_j)/h)=O_p(n^{-1})$$

请证明这个结果。

(v) 使用附录 A 引理 A.15 给出的 U 统计量 H 分解，证明 $J_n=E(J_n)+O_p(h^{1/2}(nh)^{-1}+n^{-1/2}h^4)$ + 与 h 无关项。因此，我们确实有 $J_n=E(J_n)+(\mathrm{s.o.})$（对一个给定的 h 值）。

提示：注意 $\bar{k}(\cdot)$ 也是一个非负、对称的概率密度函数，即 $\int \bar{k}(v)\mathrm{d}v = 1$，当 s 是一个奇的正整数时，$\int v^s \bar{k}(v)\mathrm{d}v = 0$。

(i)

$$E[\bar{k}_{12}] = h^{-1}E\int k\left(\frac{X_1 - x}{h}\right)k\left(\frac{X_2 - x}{h}\right)\mathrm{d}x$$

$$= h^{-1}\int\left[Ek\left(\frac{X_1 - x}{h}\right)\right]\left[Ek\left(\frac{X_2 - x}{h}\right)\right]\mathrm{d}x$$

$$= h^{-1}\int\left[Ek\left(\frac{X_1 - x}{h}\right)\right]^2\mathrm{d}x$$

$$= h\int[f(x) + 0 + (\kappa_2/2)f^{(2)}(x)h^2 + 0 + (\kappa_4/4!)f^{(4)}(x)h^4 + O(h^5)]^2\mathrm{d}x$$

$$E[k_{12}] = Ek\left(\frac{X_1 - X_2}{h}\right)$$

$$= \int k\left(\frac{x_1 - x_2}{h}\right)f(x_1)f(x_2)\mathrm{d}x_1\mathrm{d}x_2$$

$$= h\int f(x)\{f(x) + 0 + (\kappa_2/2)f^{(2)}(x)h^2 + 0 + (\kappa_4/4!)f^{(4)}(x)h^4\}\mathrm{d}x + O(h^5)$$

(ii) 注意 $\bar{k}(0) = \int k^2(v)\mathrm{d}v > 0$。

(iii) 证明 $h \to 0$ 产生的目标函数的值为 $V(h) = -\infty$。因此，$h = 0$ 最小化 $V(h)$。

52

(iv) 证明 $E\|A_n\| = E(A_n) = O(n^{-1})$，然后应用定理 A.7。

(v) 使用 U 统计量 H 分解（再次见附录 A），证明 H 分解的最后两项分别是 $O_p(n^{-1/2}h^4)$（加上与 h 无关项）和 $O_p(h^{1/2}(nh)^{-1})$ 阶的。

习题 1.7 推导式(1.27)，即证明

$$\int_{-\infty}^{x}\hat{f}(v)\mathrm{d}v = n^{-1}\sum_{i=1}^{n}G\left(\frac{x - X_i}{h}\right)$$

提示：使用 $\hat{f}(v) = (nh)^{-1}\sum_{i=1}^{n}k\left(\frac{X_i - v}{h}\right)$ 且做一个变量替换 $(x_i - v)/h = t$ 以及 $\mathrm{d}x_i = h\mathrm{d}v$。

习题 1.8

(i) 讨论核和经验累积分布函数估计量之间的关系，即 $\hat{F}(x)$ 和 $F_n(x) =$

$n^{-1}\sum_{i=1}^{n}\mathbf{1}(X_i \leqslant x)$。

(ii)讨论人们是否能在定义于式(1.27)的 $\hat{F}(x)$ 中使用 $h=0$，即人们能否在 $\hat{F}(x)$ 中让 $h\to 0$ 任意快？

(iii) $\hat{F}(x)$ 和 $F_n(x)$ 有相同的渐近分布。使用 $\hat{F}(x)$ 而不是 $F_n(x)$ 的优点是什么？你期望哪一个估计量有更小的有限样本？请解释。

习题 1.9　推导式(1.33)。

提示：记 $\mathbf{1}_i(x)=\mathbf{1}(X_i\leqslant x)$ 和 $G_{x,x_j}=G((x-X_j)/h)$，那么

$$
\begin{aligned}
E[\mathrm{CV}_F(h)] &= \frac{1}{n(n-1)^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\sum_{l\neq i}^{n}\int E\{[\mathbf{1}_i(x)-G_{x,x_j}] \\
&\quad \times [\mathbf{1}_i(x)-G_{x,x_l}]\}\mathrm{d}x \\
&= \frac{1}{n-1}\int E\{[\mathbf{1}_i(x)-G_{x,x_j}]^2\}\mathrm{d}x \\
&\quad + \frac{n-2}{n-1}\int E\{E[\mathbf{1}_i(x)-G_{x,x_j}\mid X_i]\}^2\mathrm{d}x \\
&= \mathrm{CV}_1+\mathrm{CV}_2
\end{aligned}
$$

然后证明

$$
\mathrm{CV}_1 = (n-1)^{-1}\left\{2\int F(1-F)\mathrm{d}x - C_1h + O(h^2)\right\}
$$

$$
\mathrm{CV}_2 = \left[1-\frac{1}{n-1}\right]\left\{\int F(x)(1-F(x))\mathrm{d}x + h^4\int C_2(x)^2\mathrm{d}x\right\}
$$

习题 1.10　定义一个 $q\times q$ 矩阵 A，给定其第 (t,s) 个元素为 $A_{t,s}=(\kappa_2/2)\int B_t(x)B_s(x)\mathrm{d}x$.

(i) 证明 A 是半正定的。

(ii) 证明如果 A 是正定的，那么定义于式(1.41)的 a_s^0 全部可唯一确定，且是正的和有限的。

A 是正定的一个必要条件是对所有 $s=1,\cdots,q$，有 $f_{ss}(x)$ 不是一个零函数。

提示：

(i) 注意对任意 $q\times 1$ 向量 $z=(z_1,\cdots,z_q)'$，有 $z'Az=\int\left[\sum_{s=1}^{q}B_s(x)z_s\right]^2\mathrm{d}x\geqslant 0$。

(ii) 定义 $z_s=a_s^2$，那么 $\chi_f=z'Az+\kappa^q/\sqrt{z_1,\cdots,z_q}$，并令 z_s^0 表示最小化 χ_f 的 z_s 值。很容易说明 $0<\inf_{z_1,\cdots,z_q}\chi_f<\infty$。这意味着 $z_s^0>0$ 对所有 s 成立。最后，根据 Li 和 Zhou(2005)给出的结果，z_s^0 是唯一被决定的。因此，$a_s^0=\sqrt{z_s^0}$ 是唯一被决定

的，且对所有 $s=1,\cdots,q$ 是正的和有限的。

注意 A 正定是一个充分条件。Li 和 Zhou(2005)对这个结果提供了一个更弱的充要条件。

习题 1.11 证明式(1.36)和式(1.37)。

提示：对一个多变量 Taylor 展开，我们有 $f(x_0+x)=f(x_0)+\sum_{s=1}^{q}f_s(x_0)(x_s-x_{s0})+(1/2)\sum_{s=1}^{q}\sum_{s'=1}^{q}f_{ss'}(\tilde{x})(x_s-x_{s0})(x_{s'}-x_{s'0})$，$\tilde{x}$ 位于 x_0 和 x 之间的线段上。

习题 1.12 对多变量情形，我们有

$$\mathrm{CV}_f(h_1,\cdots,h_q)=\frac{\kappa^q}{nh_1\cdots h_q}+J_n+O_p((n^2h_1,\cdots,h_q)^{-1})$$

其中 $J_n=[n(n-1)]^{-1}\sum_i\sum_{j\neq i}[\bar{K}_n(X_i,X_j)-2K_n(X_i,X_j)]$。

(i) 证明 $E(J_n)=\int\left[\sum_{s=1}^{q}B_s(x)h_s^2\right]^2\mathrm{d}x+o\left(\sum_{s=1}^{q}h_s^4\right)$，其中 B_s 的定义在 1.8 节给出。

(ii) 使用 U 统计量 H 分解来证明(忽略与 H_s 无关项)

$$J_n=E(J_n)+O_p\left(n^{-1/2}\left(\sum_{s=1}^{q}h_s^2\right)^2\right)$$
$$+O_p((h_1,\cdots,h_q)^{1/2}(nh_1,\cdots,h_q)^{-1})$$

注意(i)和(ii)一起意味着

$$\mathrm{CV}_f=\sum_{s=1}^{a}B_sh_s^4+\kappa^q(nh_1,\cdots,h_q)^{-1}+o_p(\eta_2^2+\eta_1)$$

其中 $\eta_2=\sum_{s=1}^{q}h_s^2$ 且 $\eta_1=(nh_1,\cdots,h_q)^{-1}$。

提示：使用 H 分解，证明第二和第三项的二阶矩分别是 $O(n^{-1/2}\eta_2^2)$ 和 $O((h_1\cdots h_q)\eta_1^2)$ 阶的。

习题 1.13 假定 $X\in[0,1]$ 和 $f(0)>0$，证明 $E[\hat{f}(0)]=f(0)/2+O(h)$ 使得 $\hat{f}(0)$ 是 $f(0)$ 的一个有偏估计量，甚至是渐近有偏的。

提示：$\hat{f}(0)=(nh)^{-1}\sum_{i=1}^{n}k((X_i-0)/h)$，且

$$E[\hat{f}(0)]=h^{-1}E[k(X_i/h)]=h^{-1}\int_0^1 f(x_1)k(x_1/h)\mathrm{d}x_1$$
$$=\int_0^{1/h}f(hv)k(v)\mathrm{d}v$$
$$\to f(0)\int_0^{\infty}k(v)\mathrm{d}v=f(0)/2$$

习题1.14 使用定义于式(1.43)中的修正的边界核,定义于式(1.44)中的$\hat{f}(x)$以及位于[0,1]的X的支撑,证明对在边界处的$x\in[0,h]$,我们有$E[\hat{f}(x)]=f(x)+O(h)$。明确地给出推导这个结果需要的条件。

因此,当$n\to\infty$时,$\text{bias}[\hat{f}(x)]=O(h)\to 0$,且修正的边界核恢复了当$x$在边界处时对$\hat{f}(x)$的渐近无偏性。

提示:可以记$x=\alpha h$,其中$0\leqslant\alpha\leqslant 1$。可以假定对所有$x,z\in[0,1]$;$|f(x)-f(z)|\leqslant C|x-z|$,其中$C$是一个正常数,然后令$x=\alpha h,\alpha\in[0,1]$,我们有$\hat{f}(x)=\hat{f}(\alpha h)$和

55

$$E[\hat{f}(\alpha h)]=\frac{h^{-1}\int_0^1 k\left(\frac{x_1-\alpha h}{h}\right)f(x_1)\,\mathrm{d}x_1}{\int_{-\alpha}^{\infty}k(v)\,\mathrm{d}v}$$

$$=\frac{\int_{-\alpha}^{-\alpha+1/h}k(v)f(\alpha h+vh)\,\mathrm{d}v}{\int_{-\alpha}^{\infty}k(v)\,\mathrm{d}v}=f(0)+O(h)$$

因为

$$\int_{-\alpha+1/h}^{\infty}k(v)\,\mathrm{d}v\leqslant c\int_{-\alpha+1/h}^{\infty}(1+v)^{-2}\,\mathrm{d}v=(-c)\frac{1}{1+v}\bigg|_{v=-a}^{\infty}$$

$$=-c\left[0-\frac{1}{1-\alpha+1/h}\right]=\frac{ch}{(1-\alpha)h+1}=O(h)$$

对足够大的v,给定$k(v)\leqslant c/(1+v^2)$,于是有

$$\int_{-\alpha}^{-\alpha+1/h}k(v)\,\mathrm{d}v\Big/\int_{-\alpha}^{\infty}k(v)\,\mathrm{d}v=1-\int_{-\alpha+1/h}^{\infty}k(v)\,\mathrm{d}v\Big/\int_{-\alpha}^{\infty}k(v)\,\mathrm{d}v$$

$$=1-O(h)/O(1)=1+O(h)$$

习题1.15 有一个ν阶核,对单变量x的情况证明式(1.46)和式(1.47)(即$q=1$)。

习题1.16 直观地,人们可能认为当$f(x)$是一个均匀密度,比如位于[0,1]时,能够选择一个非缩小的h对某些$x\in[0,1]$(即当$n\to\infty$时h不趋于0)来估计$f(x)$。当x是一个[0,1]的内点时,这种直觉是正确的。然而,在(或接近)[0,1]的边界,即使对均匀$f(x)$估计偏误也不会趋于0。

(i) 证明当$n\to\infty$时,如果h不趋于0,那么$\int_0^1[\hat{f}(x,h)-f(x)]^2\mathrm{d}x$将不会收敛于0,其中$f(x)$是均匀概率密度函数。

(ii) 证明当$n\to\infty$时,如果$h\to 0$,那么$\int_0^1[\hat{f}(x,h)-f(x)]^2\mathrm{d}x\to 0$,其中

$f(x)$ 是均匀概率密度函数。

以上(i)和(ii)解释了为什么当 $n \to \infty$ 时,交错鉴定选择的 h,即 $\hat{h}$ 必须收敛于0,并且为什么人们不需要 $f^{(2)}(x)$ 是非零函数这个条件。当然当 $f(x)$ 是一个均匀概率密度函数时,$\hat{h}$ 将不再有通常的阶($n^{-1/5}$),而是有一个等于 $n^{-1/3}$ 的阶,因为偏误现在是 h 阶而不是 h^2 阶。

习题 1.17 考虑来自 1.13.5 节的意大利收入数据。对 1951 年和 1998 年大小为 $n=21$ 的两个样本,假定一个潜在的正态分布,使用式(1.17)的推断经验法则计算密度估计。窗宽必须是其多少倍才能消除在 1998 年样本中出现的双峰特征?下一步,使用最小二乘交错鉴定法计算密度估计。假定这些窗宽代表最优窗宽,1998 年的窗宽必须是多大才能产生一个明显的单峰分布?最后,把你的最小二乘交错鉴定密度估计跟一个单纯的直方图比较。你的估计显得更敏感吗,即它们反映了在数据中实际出现的特征吗?

56

第2章 回 归

回归分析可能是在所有应用数据分析中被最广泛使用的工具。给定一个典型的以 X 表示的协变量向量(它通常被称作“回归元”或“解释变量”),回归方法对一个典型的以 Y 表示的因变量的期望行为建模。换句话说,回归分析被设计用于回答诸如此类问题:在交通运输部门工作的一个黑人女性大学毕业生的期望工资是多少?更进一步,应用者经常对因变量的一个或多个协变量的一个变化如何做出反应(响应)以及这种反应是否显著区别于零(显著性检验)感兴趣。我们首先简要概述参数回归,然后快速转向非参数回归的研究。 57

到目前为止最流行的参数回归模型是由下式给出的线性回归模型:

$$Y_i = \beta_0 + X_i'\beta + u_i, \quad i = 1, \cdots, n \tag{2.1}$$

其中 $X_i \in \mathbb{R}^q$,且 β 为 $q \times 1$ 维的未知参数向量,而一个更一般的非线性回归模型由下式给出:

$$Y_i = g(X_i, \beta) + u_i, \quad i = 1, \cdots, n \tag{2.2}$$

其中 $g(\cdot, \cdot)$ 有一个已知的函数形式,β 仍然是 $q \times 1$ 维的未知参数向量。例如,我们可以给出这样形式的一个模型 $g(x, \beta) = \exp(x'\beta)$。如果对包含有限个未知参数的回归模型已经设定函数形式,诸如普通最小二乘或非线性最小二乘之类的方法就能够被用于估计式(2.1)或式(2.2)中的未知参数向量 β。 58

跟用参数方法模型化一个概率密度函数时的情况一样(见第1章),在实践中真实回归函数形式几乎很少是已知的。由于参数方法要求使用者在估计前设定准确的模型参数形式,人们必须面对假定模型可能跟数据生成过程(DGP)不一致的可能性,并且在实践中人们必须处理参数回归模型可能出现的被严重误设的问题。为了得到有效的推断,人们必须正确设定的不仅仅是条件均值函数,而且也包括异方差和序列相关函数。基于

错误设定模型的估计和推断的不可忽略的后果被固定下来,包括不一致的参数估计和无效的推断。正如在第1章提到的,人们当然能够检验假定的参数模型是否正确,但拒绝某个参数模型则没有为备择模型的形式提供什么内容。也就是说,拒绝假定模型不会因此产生一个正确设定的模型。

非参数回归模型不要求应用者对潜在数据生成过程做出函数形式的假定。它并不假定人们知道待估对象的准确函数形式,而是假定该对象存在且满足某些正则条件,比如平滑性(可微性)和矩条件。然而,我们再次指出这不是没有代价的。通过对问题施加较少的结构限制,非参数方法要求更多的数据才能达到跟一个正确设定的参数模型相同的准确度。然而,如果人们怀疑一个参数模型在某种程度上被错误设定,而且手中的样本不是太小以至于使用非参数技术没有什么意义,那么应用者可以考虑非参数回归方法。

我们先考虑以下非参数回归模型

$$Y_i = g(X_i) + u_i, \quad i = 1, \cdots, n \tag{2.3}$$

我们在这里假定样本实现(Y_i, X_i)是独立同分布的,尽管我们将在第18章放松这个假定。函数形式$g(\cdot)$当然是未知的。如果$g(\cdot)$是一个平滑函数,那么我们能够用核方法非参数估计$g(\cdot)$。基于定理2.1,我们将把$g(x)$解释为给定$X = x$时Y的条件均值,即$g(x) = E[Y_i | X_i = x]$。

59

定理 2.1 令$\mathcal{G}$表示具有有限二阶矩的Borel可测(或连续)函数族(见附录A对Borel可测函数的定义)。假定$g(x) \equiv E(Y|X = x)$属于$\mathcal{G}$,并且$E(Y^2)$是有限的。那么,在如下均方误差(MSE)的意义上:

$$E\{[Y - r(X)]^2\} \geqslant E\{[Y - E(Y|X)]^2\}, \quad \text{所有 } r(\cdot) \in \mathcal{G}$$

或者,等价地,

$$\min_{\{r(\cdot) \in \mathcal{G}\}} E\{[Y - r(X)]^2\} = E\{[Y - E(Y|X)]^2\}$$

即$E(Y|X)$是给定X时Y的最优预测值。

证明:首先我们观察到$g(x) = E(Y|X = x)$是一个x的函数。接下来令$f_{y,x}(x,y)$、$f(x)$和$f_{y|x}(y|x)$分别表示(Y,X)的联合概率密度函数、X的边际概率密度函数,以及$Y|X$的条件概率密度函数。从$f_{y|x}(y|x) = f_{y,x}(x,y)/f(x)$,我们有

$$E(Y|X = x) = \int y f_{y|x}(y|x)\mathrm{d}y = \frac{\int y f_{y,x}(x,y)\mathrm{d}y}{f(x)} \stackrel{\text{def}}{=} g(x) \tag{2.4}$$

这显然是一个x的函数。现在,对任意一个函数$r(x)$,我们有

$$\begin{aligned} E\{[Y - r(X)]^2\} &= E\{[Y - E(Y|X) + E(Y|X) - r(X)]^2\} \\ &= E\{[Y - E(Y|X)]^2\} + E\{[E(Y|X) - r(X)]^2\} \\ &\quad + 2E\{[(Y - E(Y|X))(E(Y|X) - r(X))]\} \end{aligned}$$

$$= E\{[Y - E(Y|X)]^2\} + E\{[E(Y|X) - r(X)]^2\}$$
$$\geqslant E\{[Y - E(Y|X)]^2\}$$

其中我们根据重复期望定律(见附录 A,引理 A.11)使用如下事实,

$$E\{[Y - E(Y|X)][E(Y|X) - r(X)]\}$$
$$\equiv E\{[Y - E(Y|X)][g(X) - r(X)]\}$$
$$\equiv E\{[g(X) - r(X)] \times E[Y - E(Y|X)|X]\} = 0$$

□

定理 2.1 说明在所有 Borel 可测(或连续)函数 $r(x)$ 族中,$E(Y|X)$ 在均方误差最小化的意义上是利用 X 预测 Y 的最优函数。于是我们将把式(2.3)中的 60 $g(x)$ 解释为 $E(Y|x)$。考虑到式(2.4),很明显一旦人们知道如何使用在第1章中提出的方法估计 $f_{y,x}(x,y)$ 和 $f(x)$,人们就能够轻易得到 $g(x) = E(Y|x)$ 的一个估计量。这直接引出了"局部常数"核估计量,它最初由 Nadaraya(1965)和 Watson(1964)提出,经常被简单地称为"Nadaraya-Watson 核估计量"。

2.1 局部常数核估计

在第1章我们详细研究了 $f_{y,x}(x,y)$ 和 $f(x)$ 的估计,那么为估计一个条件均值,仅需要额外的一步,即关于 y 积分得到 $\int y f_{y,x}(x,y)\mathrm{d}y$。注意我们可以通过使用未知概率密度函数 $f_{y,x}(x,y)$ 的核估计代替它产生 $\int y\hat{f}_{y,x}(x,y)\mathrm{d}y$ 从而估计 $\int y f_{y,x}(x,y)\mathrm{d}y$,其中

$$\hat{f}_{y,x}(x,y) = \frac{1}{nh_0h_1\cdots h_q}\sum_{i=1}^{n}K\left(\frac{X_i - x}{h}\right)k\left(\frac{y - Y_i}{h_0}\right)$$

其中 $K\left(\frac{X_i - x}{h}\right) = k\left(\frac{X_{i1} - x_1}{h_1}\right) \times \cdots \times k\left(\frac{X_{iq} - x_q}{h_q}\right)$ 且 h_0 是跟 Y 有关的平滑参数。因此,我们有

$$\int y\hat{f}_{y,x}(x,y)\mathrm{d}y = \frac{1}{nh_0h_1\cdots h_q}\sum_{i=1}^{n}K\left(\frac{X_i - x}{h}\right)\int yk\left(\frac{y - Y_i}{h_0}\right)\mathrm{d}y$$
$$= \frac{1}{nh_0h_1\cdots h_q}\sum_{i=1}^{n}K\left(\frac{X_i - x}{h}\right)\int (Y_i + h_0 v)k(v)h_0\mathrm{d}v$$

(变量变换: $(y - Y_i)/h_0 = v$)

$$= \frac{1}{nh_1\cdots h_q}\sum_{i=1}^{n}K\left(\frac{X_i - x}{h}\right)Y_i \tag{2.5}$$

其中,我们使用了 $\int k(v)\mathrm{d}v = 1$ 和 $\int vk(v)\mathrm{d}v = 0$。

基于式(2.4)和式(2.5),我们通过下式估计 $E(Y|x) \equiv g(x)$

$$\hat{g}(x) = \frac{\int y\hat{f}_{y,x}(x,y)\mathrm{d}y}{f(x)} = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} \tag{2.6}$$

61 这正好是 Y_i 的一个加权平均,因为我们能重写式(2.6)为

$$\hat{g}(x) = \sum_{i=1}^{n} Y_i w_i$$

其中 $w_i = K\left(\frac{X_i - x}{h}\right) \Big/ \sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)$ 是施加于 Y_i 的权重。注意权重非负并且其和为 1。

沿着第 1 章中给出的分析思路,我们能够证明

$$\hat{g}(x) - g(x) = O_p\left(\sum_{s=1}^{q} h_s^2 + (nh_1 \cdots h_q)^{-1/2}\right)$$

它的证明类似于下式的证明

$$\hat{f}(x) - f(x) = O_p\left(\sum_{s=1}^{q} h_s^2 + (nh_1 \cdots h_q)^{-1/2}\right)$$

它在第 1 章 1.6 节出现过。证明 $\hat{g}(x)$ 是一致估计的一个简单方法是对分子和分母分别处理。首先我们记

$$\hat{g}(x) - g(x) = \frac{(\hat{g}(x) - g(x))\hat{f}(x)}{\hat{f}(x)} \equiv \frac{\hat{m}(x)}{\hat{f}(x)} \tag{2.7}$$

其中 $\hat{m}(x) = (\hat{g}(x) - g(x))\hat{f}(x)$。使用 $Y_i = g(X_i) + u_i$,我们有

$$\hat{m}(x) = \hat{m}_1(x) + \hat{m}_2(x)$$

其中

$$\hat{m}_1(x) = (nh_1 \cdots h_q)^{-1} \sum_{i=1}^{n} (g(X_i) - g(x)) K\left(\frac{X_i - x}{h}\right)$$

且

$$\hat{m}_2(x) = (nh_1 \cdots h_q)^{-1} \sum_{i=1}^{n} u_i K\left(\frac{X_i - x}{h}\right)$$

使用记号 $r(x + hv) = r(x_1 + h_1 v_1, \cdots, x_q + h_q v_q)$ 和使用在式(1.36)的推导中同样

62 的论证,我们能轻易证明

$$E[\hat{m}_1(x)] = (h_1 \cdots h_q)^{-1} \int f(x_1)[g(x_1) - g(x)] K\left(\frac{x_1 - x}{h}\right) \mathrm{d}x_1$$

$$= \int f(x+hv)[g(x+hv)-g(x)]K(v)\mathrm{d}v$$

$$= \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2\{[2f_s(x)g_s(x)+f(x)g_{ss}(x)]\} + O\left(\sum_{s=1}^{q} h_s^3\right)$$

$$\equiv f(x)\sum_{s=1}^{q} h_s^2 B_s(x) + O\left(\sum_{s=1}^{q} h_s^3\right) \tag{2.8}$$

其中 $\kappa_2 = \int v^2 k(v)\mathrm{d}v$，$B_s(x) = \frac{\kappa_2}{2}\{2f_s(x)g_s(x)+f(x)g_{ss}(x)\}/f(x)$，且 $r_s(x)$ 和 $r_{ss}(x)$ 分别是 $r(x)$ 关于 x_s 的一阶和二阶导数（$r=g$ 或 $r=f$）。同样，

$$\begin{aligned}
\mathrm{var}(\hat{m}_1(x)) &= (nh_1^2\cdots h_q^2)^{-1}\mathrm{var}\left[(g(X_1)-g(x))K\left(\frac{X_1-x}{h}\right)\right] \\
&= (nh_1^2\cdots h_q^2)^{-1}\left(E\left\{\left[(g(X_1)-g(x))K\left(\frac{X_1-x}{h}\right)\right]^2\right\}\right. \\
&\quad \left. - \left\{E\left[(g(X_1)-g(x))K\left(\frac{X_1-x}{h}\right)\right]\right\}^2\right) \\
&= (nh_1^2\cdots h_q^2)^{-1}\left((h_1\cdots h_q)\int f(x+hv)\right. \\
&\quad \left.\times[g(x+hv)-g(x)]^2K(v)^2\mathrm{d}v - (\mathrm{s.o.})\right) \\
&= O\left((nh_1\cdots h_q)^{-1}\sum_{s=1}^{q} h_s^2\right)
\end{aligned} \tag{2.9}$$

其中(s. o.)代表低阶项。

式(2.8)和式(2.9)导出

$$\hat{m}_1(x) = f(x)\sum_{s=1}^{q} B_s(x)h_s^2 + O_p(\eta_2^{3/2} + \eta_2^{1/2}\eta_1^{1/2}) \tag{2.10}$$

其中 $\eta_1 = (nh_1\cdots h_q)^{-1}$ 且 $\eta_2 = \sum_{s=1}^{q} h_s^2$（见习题 2.1）。接下来，我们注意到 $E[\hat{m}_2(x)]=0$，那么

63

$$\begin{aligned}
E\{[\hat{m}_2(x)]^2\} &= (nh_1^2\cdots h_q^2)^{-1}E\left[u_i^2K^2\left(\frac{X_i-x}{h}\right)\right] \\
&= (nh_1^2\cdots h_q^2)^{-1}E\left[\sigma^2(X_i)K^2\left(\frac{X_i-x}{h}\right)\right] \\
&= (nh_1\cdots h_q)^{-1}\left\{f(x)\sigma^2(x)\int K(v)^2\mathrm{d}v + O\left(\sum_{s=1}^{q} h_s^2\right)\right\} \\
&\equiv (nh_1\cdots h_q)^{-1}\Omega(x) + O\left((nh_1\cdots h_q)^{-1}\sum_{s=1}^{q} h_s^2\right)
\end{aligned} \tag{2.11}$$

其中 $\Omega(x)=\kappa^q f(x)\sigma^2(x)$，$\kappa=\int k(v)^2\mathrm{d}v$ 且 $\sigma^2(x)=E(u_i^2|X_i=x)$。

由式(2.10)和式(2.11)可得

$$\hat{m}(x)=\hat{m}_1(x)+\hat{m}_2(x)=O_p(\eta_2+\eta_1^{1/2}) \tag{2.12}$$

因此，给定 $\hat{g}(x)-g(x)=\hat{m}(x)/\hat{f}(x)$ 且 $\hat{f}(x)=f(x)+o_p(1)$，我们可证明如下的结果（假定 $f(x)>0$）：

$$\begin{aligned}\hat{g}(x)-g(x)&=O_p\left(\frac{\hat{m}(x)}{[f(x)+o_p(1)]}\right)\\&=O_p\left(\sum_{s=1}^{q}h_s^2+(nh_1\cdots h_q)^{-1/2}\right)\end{aligned} \tag{2.13}$$

从式(2.13)很容易看出，如果每个窗宽(h_s)有同样大小的阶，那么，最小化 MSE[$\hat{g}(x)$]的 h_s 的最优选择是 $h_s\sim n^{-1/(q+4)}$，并且相应的均方误差是 $O(n^{-4/(q+4)})$阶的。

在这个阶段使用 Liapunov 中心极限定理也能证明 $\hat{g}(x)$的渐近正态性。

定理 2.2 在 x 是一个内点，$g(x)$和 $f(x)$是三次连续可微的，且 $f(x)>0$ 的假定下，那么当 $n\to\infty$ 时，$h_s\to 0$（对所有 s），$nh_1\cdots h_q\to\infty$ 且 $(nh_1\cdots h_q)\sum_{s=1}^{q}h_s^6\to 0$，我们有

$$\sqrt{nh_1\cdots h_q}\left(\hat{g}(x)-g(x)-\sum_{s=1}^{q}h_s^2B_s(x)\right)\xrightarrow{d}N(0,\kappa^q\sigma^2(x)/f(x)) \tag{2.14}$$

其中 $B_s(x)$是直接根据式(2.8)定义的。

证明：$(nh_1\cdots h_q)^{1/2}\hat{m}_2(x)$有均值零且式(2.11)证明了它有渐近方差$\Omega(x)$。使用 Liapunov 中心极限定理，容易证明

$$\sqrt{nh_1\cdots h_q}\hat{m}_2(x)\xrightarrow{d}N(0,\Omega(x)) \tag{2.15}$$

结合式(2.10)和式(2.15)，我们有

$$\begin{aligned}\sqrt{nh_1\cdots h_q}\left(\hat{m}(x)-\sum_{s=1}^{q}h_s^2B_s(x)\hat{f}(x)\right)&=\sqrt{nh_1\cdots h_q}\hat{m}_2(x)+o_p(1)\\&\xrightarrow{d}N(0,\Omega(x))\end{aligned} \tag{2.16}$$

因为 $(nh_1\cdots h_q)^{1/2}\eta_2^{3/2}\sim(nh_1\cdots h_q)^{1/2}\sum_{s=1}^{q}h_s^3=o(1)$。

从式(2.16)，并且注意 $\hat{f}(x)=f(x)+o_p(1)$，我们立刻得到

$$\sqrt{nh_1\cdots h_q}\left(\hat{g}(x)-g(x)-\sum_{s=1}^{q}h_s^2B_s(x)\right)$$

$$\equiv\frac{\sqrt{nh_1\cdots h_q}\left(\hat{g}(x)-g(x)-\sum_{s=1}^{q}h_s^2B_s(x)\right)\hat{f}(x)}{\hat{f}(x)}$$

$$=\frac{\sqrt{nh_1\cdots h_q}\left(\hat{m}(x)-\sum_{s=1}^{q}h_s^2B_s(x)\hat{f}(x)\right)}{\hat{f}(x)}$$

$$\xrightarrow{d}\frac{1}{f(x)}N(0,\Omega(x))=N(0,\kappa^q\sigma^2(x)/f(x))\tag{2.17}$$

这就完成了定理2.2的证明。 □

2.1.1 局部常数核估计量潜在的直觉

为了掌握非参数核回归方法,人们不仅应该掌握比如推导收敛速度等基本技能,还应明白为什么核方法是那样运作的。

幸运的是,在 $\hat{g}(x)\xrightarrow{p}g(x)$ 背后的直觉是相当简单的。基本思想是 $g(x)$ 仅仅是"局部平均"(local average),在一个使用了均匀核仅有一个回归元($X_i\in\mathbb{R}$) 65 的简单情形中,这也许能被最好地解释。在这种情况下($h_1=h$)我们看到

$$\hat{g}(x)=\frac{\sum_{i=1}^{n}Y_iK\left(\frac{X_i-x}{h}\right)}{\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)}$$

$$=\frac{\sum_{|X_i-x|\leqslant h}Y_i(1/2)}{\sum_{|X_i-x|\leqslant h}(1/2)}$$

$$=\frac{\sum_{|X_i-x|\leqslant h}[g(X_i)+u_i]}{\sum_{|X_i-x|\leqslant h}1}$$

$$=\{\text{当}\ |X_i-x\leqslant h\ \text{时},g(X_i)\ \text{的平均}+u_i\ \text{的平均}\}$$

$$\xrightarrow{p}g(x)+0=g(x)$$

因为在假定 $g(x)$ 在 x 存在一个有界导数时,对所有 $|X_i-x|\leqslant h$,有 $|g(X_i)-g(x)|=O(h)=o(1)$。事实上,给定核函数的对称性,我们有 $n^{-1}\sum_{|X_i-x|\leqslant h}(g(X_i)-g(x))=O_p(h^2)$。$nh\to\infty$ 的条件保证了渐近性,无限多的观测值出现在每一

个长度为 h 的区间(或 $2h$,即我们运用与点 $X=x$ 距离在 h 内的两侧点计算 $g(x)$ 的估计值)。这是因为区间的数目是 $O(1/h)$ 阶的。我们将期望在长度为 h 的区间内平均有 $n/O(1/h)=O(nh)$ 个观测值。Azzalini 和 Bowman(1997)把 nh 称作"局部样本规模"(local sample size),这是一种考虑问题的好方法,因为这是跟在一个固定点 x 估计回归函数有关的样本。因此一致性要求"局部样本"的规模必须随其总样本大小($n\to\infty$)而增加($nh\to\infty$),而同时区间宽度在极限($h\to0$)时缩小为零。因为 $n\to\infty$ 时 $nh\to\infty$,根据大数定律的一个结论,在每个区间,u_i 的均值收敛于它们的总体样本均值 $E(u_i)=0$(在概率上)。因此,我们看到,对均匀核估计,$\hat{g}(x)$ 使用了 Y_i 的一个简单局部平均(基于与 x 靠近的那些 X_i)来估计 $g(x)$。

概括起来,我们通过将因变量的一些数值进行局部平均来估计条件均值函数,而这些数值所对应的回归元的取值是相当接近的。用来构造均值的局部信息量被一个窗宽控制。通过控制用于构造估计量的局部信息量(局部样本规模)和允许当样本规模变大(同时邻域缩小)时局部均值的大小能提供更多的信息,我们也能够保证在标准的正则条件下估计量是一致的。

66

估计量 $\hat{g}(x)$ 常被称作 $g(x)$ 的局部常数估计量,尽管人们可以仅仅简单使用局部线性/多项式方法来估计 $g(x)$,而对局部多项式方法的一种更有深度处理,我们推荐读者参考 Fan 和 Gijbels(1996)的优秀著作。局部线性估计量也自动提供一个响应估计量,即 $g(x)$ 的导数,而一个 p 阶局部多项式方法可以估计直到 p 阶的导数。我们将在下面的 2.4 节讨论局部线性和局部多项式估计方法,下面我们首先讨论非参数常数回归的基本内容,即窗宽选择。

2.2 局部常数窗宽选择

在这一节我们讨论通过在式(2.6)给出的局部常数估计量来估计未知回归函数 $g(\cdot)$ 的多种平滑参数的选择方法。我们讨论三种不同的平滑参数选择方法:(1) 经验法则和插入法,(2) 最小二乘交错鉴定法,以及(3) 一个修正的 AIC 程序。

2.2.1 经验法则和插入法

当使用一个二阶核时,可以证明最优平滑参数的阶应该是 $O(n^{-1/(4+q)})$。一个流行的经验法则程序是通过 $c_sX_{s,sd}n^{-1/(4+q)}$ 来选择 $h_s(s=1,\cdots,q)$,其中 c_s 是一个常数,$X_{s,sd}$ 是 $\{X_{is}\}_{i=1}^n$ 的样本标准差。在应用中,c_s 经常被选作 1 或者某个其他接近 1 的常数。经验法则背后的论证与在第 1 章讨论的所谓密度估计

的正态推断规则是相关的。经验法则对应用者是有吸引力的,因为从计算上来说它们简单。然而,这种方法的一个缺陷是它们把 x 的所有元素(协变量)对称地对待。在实践中,一个回归函数 $g(x)$ 可能关于一个元素缓慢变化,比如 x_1,但关于另一个元素迅速变化,比如 x_2。在这种情况下,人们应该对 x_1 使用一个相对较大的平滑参数而对 x_2 使用一个相对较小的平滑参数。很明显经验法则缺少这种灵活性。

67

一种可供选择的方法是所谓的插入法(plug-in method),它通常基于最小化一个形式为 $\int E[\hat{g}(x)-g(x)]^2\nu(x)\mathrm{d}x$ 的"加权积分均方误差"(WIMSE),其中期望是关于随机样本 $\{X_i,Y_i\}_{i=1}^n$ 取的,并且 $\nu(x)$ 是一个保证加权积分均方误差是渐近有限的非负权重函数。

在式(2.14)中推导出了 $\hat{g}(x)$ 的偏误和方差首项。因此,$\int E[\hat{g}(x)-g(x)]^2\nu(x)\mathrm{d}x$ 的首项由下式给出

$$\begin{aligned}\mathrm{WIMSE} &= \int\left\{\left[\sum_{s=1}^{q}h_s^2B_s(x)\right]^2+\frac{\kappa^q}{nh_1\cdots h_q}\frac{\sigma^2(x)}{f(x)}\right\}\nu(x)\mathrm{d}x \\ &= O\left(\left(\sum_{s=1}^{q}h_s^2\right)^2+(nh_1\cdots h_q)^{-1}\right)\end{aligned}\tag{2.18}$$

注意此处使用一个权重函数是重要的,因为对许多熟悉的密度函数(如正态密度函数),$\int\sigma^2(x)f(x)^{-1}\mathrm{d}x$ 不存在。

通过 $h_s=a_sn^{-1/(q+4)}$ 定义 a_s,那么式(2.18)变为

$$\mathrm{WIMSE}=n^{-4/(q+4)}\chi_\nu(a_1,\cdots,a_q)\tag{2.19}$$

其中

$$\chi_\nu(a_1,\cdots,a_q)=\int\left\{\left[\sum_{s=1}^{q}a_s^2B_s(x)\right]^2+\frac{\kappa^q}{a_1\cdots a_q}\frac{\sigma^2(x)}{f(x)}\right\}\nu(x)\mathrm{d}x\tag{2.20}$$

我们令 $a_1^0,\cdots,a_q^0$ 代表最小化 $\chi_\nu(a_1,\cdots,a_q)$ 的 $a_1,\cdots,a_q$ 值。如果其中一个 a_s^0 是0,我们对某 $t\neq s$ 必有 $a_t^0=\infty$。这意味着 $h_t=\infty$。如果 $h_t=\infty$,我们容易看出 $k((X_{it}-x_t)/h_t)=k(0)$ 成为一个常数,并且这个常数在 $\hat{g}(x)$ 的分子和分母中消掉了,以至于 $\hat{g}(x)$ 跟 x_t 无关(如果 $h_t=\infty$)。在这一节,我们假定 x 的所有元素是相关的回归元,因此我们排除 $a_t^0(h_t)$ 是无穷大的情况并且假定①

68

每个 a_t^0 是唯一被定义的、正的且有限的。 (2.21)

---

① 我们将在2.2.4节讨论"无关回归元"(irrelevant regressor)的情况,那里我们将看到一个无关回归元的最优平滑参数应该是无穷大的。

令 h_s^0 的代表最小化式(2.18)的平滑参数,我们有

$$h_s^0 = n^{-1/(q+4)} a_s^0,\ s = 1,\cdots,q \tag{2.22}$$

式(2.22)意味着 $h_s^0 = O(n^{-1/(q+4)})$。我们观察到 a_s^0 依赖于未知函数 $g(\cdot)$、$f(\cdot)$和它们的导数(因为 $B_s(x)$依赖于这些函数)。当 $1 \leqslant q \leqslant 2$ 时,得到 a_s^0 的显式解是相当直接的(见习题2.2)。然而,对一般高维情况,关于 a_s^0 的显式解不存在,即使 a_s 是很好定义的(再见习题2.2)。

当一个关于 a_s^0 的显式解存在时,人们可以得到 a_s^0 的一个一致估计量。令 $\hat{a}_s^0$ 表示 a_s^0 的一个一致估计量,人们能够通过 $\hat{h}_s^0 = \hat{a}_s^0 n^{-1/(q+4)}$ $(s=1,\cdots,q)$ 估计 h_s^0。这种方法被称作插入法选择平滑参数。即使关于 a_s^0 的显式解不存在,从理论上人们仍然能获得 a_s^0 的一致估计。例如,人们能用某些一致估计量,比如 $\hat{B}_s(x)$的估计来代替 $B_s(x)$,计算包括χ_ν 的积分,然后关于 $a_1,\cdots,a_q$ 在数值上最小化χ_ν 来得到 $\hat{a}_1^0,\cdots,\hat{a}_q^0$。这样一个过程可能计算量很大。如同我们在2.2.2节讨论的,在这种情况下,人们还可以用另外一种交错鉴定程序来最优地选择平滑参数。

即使存在一个关于 a_s^0 的显式解,人们仍然需要 $B_s(x)$和 $\sigma^2(x)$的初始非参数估计,这要求人们轮流选择一个初始的或"试验的"平滑参数来估计这些未知量。初始的平滑参数可以通过经验法则方法比如以上讨论的 $h_s = x_{s,sd} n^{-1/(q+4)}$ $(s=1,\cdots,q)$来选择。然而,如果初始选择的平滑参数离它们的最优值 h_s^0 较远,第二步插入的 $\hat{h}_s^0$ 可能也会离 h_s^0 比较远,如同 Loader(1999)所讨论的。最后,当条件(2.21)不成立时,插入法也不再被很好地定义。我们在下一节证明最小二乘交错鉴定法是不受以上提到的各个问题影响的。可能由于这些原因,插入法经常被用于多变量非参数的实际应用中。

69

2.2.2 最小二乘交错鉴定法

在这一节我们介绍一个完全数据驱动的平滑参数选择方法,被称为"局部常数最小二乘交错鉴定法"(local constant least squares cross-validation)。我们选择 $h_1,\cdots,h_q$ 来最小化目标函数

$$\mathrm{CV}_{lc}(h_1,\cdots h_q) = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{g}_{-i}(X_i))^2 M(X_i) \tag{2.23}$$

其中 $\hat{g}_{-i}(X_i) = \sum_{l\neq i}^{n} Y_l K((X_i - X_l)/h) \Big/ \sum_{l\neq i}^{n} K((X_i - X_l)/h)$ 是 $g(X_i)$的去一(leave-one-out)核估计量,且 $0 \leqslant M(\cdot) \leqslant 1$ 是一个权重函数,它用于避免由被零除引起的困难或者由于边界效应引起的缓慢收敛速度(定义于1.10节)。

在 2.7 节我们证明

$$\mathrm{CV}_{lc}(h_1,\cdots,h_q) = n^{-4/(q+4)}\chi(a_1,\cdots,a_q) + o(n^{-4/(q+4)}) + (\text{与 } h \text{ 不相关的项}) \tag{2.24}$$

其中 a_s 是由 $h_s = a_s n^{-1/(q+4)}(s=1,\cdots,q)$ 定义的，且

$$\chi(a_1,\cdots,a_q) = \int\left\{\sum_{s=1}^{q} B_s(x)a_s^2\right\}^2 f(x)M(x)\mathrm{d}x + \frac{\kappa^q}{a_1\cdots a_q}\int\sigma^2(x)M(x)\mathrm{d}x \tag{2.25}$$

比较在式(2.20)给出的 χ_ν 和式(2.25)的 χ，我们注意到，如果在式(2.20)中取 $\nu(x)=f(x)M(x)$，那么 $\chi_\nu=\chi$。这一点都不奇怪，因为 χ_ν 跟下式的首项相关：

$$\int E[\hat{g}(x) - g(x)]^2\nu(x)\mathrm{d}x \tag{2.26}$$

而 χ 跟 $E[\mathrm{CV}_{lc,0}]$ 的首项相关，其中 $\mathrm{CV}_{lc,0} = n^{-1}\sum_i(g_i-\hat{g}_i)M_i$ 是 CV_{lc} 的首项，$g_i=g(X_i)$ 且 $\hat{g}_i = g_{-i}(X_i)$。因此，我们有

70

$$\begin{aligned} E[\mathrm{CV}_{lc,0}] &= E\{[\hat{g}_i - g_i]^2M_i\} = E\{E[(\hat{g}_i-g_i)^2M_i \mid X_i]\} \\ &= \int E[(\hat{g}_i-g_i)^2M_i \mid X_i = x]f(x)\mathrm{d}x \\ &= \int E[\hat{g}(x)-g(x)]^2M(x)f(x)\mathrm{d}x \end{aligned} \tag{2.27}$$

其中最后的等式使用了 $\hat{g}_i$ 是去一估计量因此根据假定 $\{X_j,Y_j\}_{j=1,j\neq i}^{n}$ 独立于 X_i 的事实。比较式(2.26)和式(2.27)，我们发现，如果 $\nu(x)=f(x)M(x)$，那么，它们是相同的。

沿着 2.2.1 节提出的分析思路，令 $a_1^0,\cdots,a_q^0$ 表示在它们每一个是非负的约束条件下最小化 χ 的 $a_1,\cdots,a_q$ 值。如同式(2.21)的情况，我们假定

每个 a_s^0 是唯一被定义的、正的和有限的。 (2.28)

令 h_s^0 代表最小化 $E[\mathrm{CV}_{lc,0}(h)]$ 首项的 h_s 值$(s=1,\cdots,q)$，那么

$$h_s^0 = a_s^0 n^{-1/(q+4)},\quad s=1,\cdots,q \tag{2.29}$$

如果我们令 $\hat{h}_s(s=1,\cdots,q)$ 表示由交错鉴定法选出的 h_s 值，那么，由于 $\mathrm{CV}_{lc}(h)=E[\mathrm{CV}_{lc,0}(h)] + (\mathrm{s.o.})$ + 与 h 不相关的项，我们能证明 $\hat{h}_s = h_s^0 + o_p(h_s^0)$。因此，我们得到本节的主要结果。

定理 2.3 在 2.7 节和式(2.28)给出的条件下，

$$n^{1/(q+4)}\hat{h}_s \to a_s^0(\text{依概率收敛}),\quad 1\leqslant s\leqslant q$$

且

$$n^{4/(q+4)}\{CV_{lc}(\hat{h}_1,\cdots,\hat{h}_q)-n^{-1}\sum_i u_i^2 M(X_i)\}$$
$$\to \inf_{a_1,\cdots,a_q}\chi(a_1,\cdots,a_q)\text{(依概率收敛)}$$

对应用研究者而言，交错鉴定法是相当有吸引力的。因为他们不必知道最优平滑参数的显性表达式；相反，他们所需要的是最小化定义于式(2.23)的目标函数 CV_{lc}，这可以通过任何标准的数值最优方法求解。即使 $h_s^0(a_s^0)$ 没有显式解，定理2.3说明通过交错鉴定法选择的平滑参数($\hat{h}_s$)渐近等价于确定性最优平滑参数 h_s^0，因而是最优的。

71

定理2.3的一个严格证明必须涉及证明低阶项在$(h_1,\cdots,h_q)\in H_n$中是一致无穷小的，其中$(h_1,\cdots,h_q)\in H_n$定义于2.7节，并且在$x\in \operatorname{supp}M$($M$的支撑；细节见Hall等(2006))。然而，当证明某些低阶项的一致性时，通过牺牲严密性和省略一些细节，定理2.3从直觉上是容易理解的。方程(2.29)意味着$n^{1/(4+q)}h_s^0\to a_s^0$，而$CV_{lc}=E[CV_{lc,0}]+(\text{s. o.})$的事实意味着$n^{1/(4+q)}\hat{h}_s=n^{1/(4+q)}h_s^0+(\text{s. o.})=a_s^0+o_p(1)\to a_s^0$(依概率收敛)。

假定式(2.28)是定理2.3成立所需要的最重要的条件，我们现在详细地讨论这个条件。定义$z=(a_1^2,\cdots,a_q^2)'$，$C_0=\kappa^q\int\sigma^2(x)M(x)\mathrm{d}x$，且令$A$代表一个$q\times q$矩阵，其第$(t,s)$个元素为

$$\int B_t(x)B_s(x)f(x)M(x)\mathrm{d}x$$

我们把式(2.25)写为

$$\chi_z(z_1,\cdots,z_q)\overset{\text{def}}{=}z'Az+\frac{C_0}{\sqrt{z_1\cdots z_q}}\tag{2.30}$$

令$z_1^0,\cdots,z_q^0$表示在它们都是非负的约束条件下最小化χ_z的$z_1,\cdots,z_q$值。我们想找到保证$z_1^0,\cdots,z_q^0$全部非负且有限的条件(这等价于式(2.28))。可以证明①

如果A是正定的，那么$z_1^0,\cdots,z_q^0$是全部正且有限的。 (2.31)

注意，由下式可知A是一个正半定矩阵

$$z'Az=\int\Big[\sum_{s=1}^q B_s(x)z_s\Big]^2 f(x)M(x)\mathrm{d}x\geqslant 0$$

对所有的z成立。条件(2.31)指出A正定是$z_1^0,\cdots,z_q^0$有限的充分条件。这里我们对式(2.31)给出一个直观的证明，并且把一个更严格的证明放在习题

72

① 注意式(2.31)是一个充分条件。一个对$z_1^0,\cdots,z_q^0$全部为正且有限的(弱)充要条件能够在Li和Zhou(2005)中找到。

2.2。首先,注意 A 正定意味着对所有 s 满足 $z_s^0<\infty$,否则我们将有 $\chi_z=\infty$。其次,对所有 s 我们有 $z_s^0>0$,否则我们将有 $C_0/(z_1^0,\cdots,z_q^0)^{1/2}=\infty$。因此,我们必须有 $0<z_s^0<\infty$,对所有 $s=1,\cdots,q$ 成立。

定理2.3仅仅包括了其中 x 的所有元素都是相关的情况。如同我们在2.2.1节所讨论的,当 $q=2$ 以及对所有 $(x_1,x_2)\in\mathbb{R}^2$,$g(x_1,x_2)=\theta(x_1)$ 时(x_2 是一个无关回归元),那么式(2.28)不成立。然而,我们能够证明交错鉴定法仍然产生最优平滑参数选择,在这种情况下最优平滑应该有性质 $\hat{h}_1=O_p(n^{-1/5})$ 和 $\hat{h}_2\to\infty$,如同我们将在2.2.4节证明的。

2.2.3 AIC_c

Hurvich, Simonoff 和 Tsai(1998)提出一种不同的窗宽选择方法,有给人印象深刻的有限样本性质。他们的方法是基于一种改进的 Akaike 信息准则(AIC,见 Akaike(1974))。Hurvich 等的准则为非参数模型提供了一个期望 Kullback-Leibler 信息的近似无偏估计。Akaike 的信息准则最初是为参数模型设计的,而 Hurvich 等的方法对可以写为结果的线性组合的估计量是有效的,因此直接被应用于广泛的非参数估计量。他们的准则由下式给出

$$\text{AIC}_c=\ln(\hat{\sigma}^2)+\frac{1+\text{tr}(H)/n}{1-\{\text{tr}(H)+2\}/n}$$

其中

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}\{Y_i-\hat{g}(X_i)\}^2=Y'(I-H)'(I-H)Y/n$$

$\hat{g}(X_i)$是一个非参数估计量且 H 是一个 $n\times n$ 的加权函数(即核权矩阵),其第(i,j)个元素由 $H_{ij}=K_{h,ij}\Big/\sum_{l=1}^{n}K_{h,il}$,$K_{h,ij}=\prod_{s=1}^{q}h_s^{-1}k((X_{is}-X_{js})/h_s)$ 给出。

通过模拟实验,Hurvich 等(1998)证明 AIC_c的窗宽选择值跟插入法(当它可用时)和许多扩展的交错鉴定法(Craven 和 Wahba(1979))相比表现得很好。 73 然而 AIC_c选择值的最优性没有可利用的严格的理论结果,Hurvich 等推测 AIC_c 选择值跟2.2.2节提出的最小二乘交错鉴定方法享有相同的渐近最优性质。的确,在 Li 和 Racine(2004a)中报告的模拟结果证明了这一点,对小样本,AIC_c 倾向于比最小二乘交错鉴定方法表现得更好,而对大样本这两种方法没有明显的差别。

2.2.4 无关回归元出现的情况

我们现在考虑其中某些回归元不相关的情况(即允许这种可能性)。不失一般性,我们假定仅仅 x 的前 q 个元素在如下定义的意义上是相关的。对整数

$1 \leqslant q_1 \leqslant q, 0 \leqslant q_2 \leqslant q-1$ 满足 $q_1+q_2=q$，令 $\bar{X}$ 由 X 的前 q_1 个元素构成且令 $\tilde{X}=X/\bar{X}$ 表示 X 的剩余元素（即它的补集）。我们假定

$$(Y, \bar{X}) \text{是独立于 } \tilde{X} \text{ 的。} \tag{2.32}$$

式(2.32)的一个结论是几乎确定有 $E[Y|\bar{X},\tilde{X}]=E[Y|\bar{X}]$，因此 $\bar{X}$ 仅仅包含相关回归元而 $\tilde{X}$ 仅仅包含不相关的回归元。然而，很重要的是，我们不用假定这是先验上已知的。在实践中我们用非参数方法估计 $E[Y_i|X_i]$，而不是 $E[Y_i|\bar{X}_i]$。我们将证明如果最小二乘交错鉴定法被用于选择平滑参数，那么渐近的无关变量可以被自动平滑掉。为了更具体地知道“平滑掉无关变量”是什么意思，注意

$$\hat{g}(x)=\frac{\sum_{i=1}^{n} Y_i \prod_{s=1}^{q_1} k\left(\frac{X_{is}-x_s}{h_s}\right) \prod_{s=q_1+1}^{q} k\left(\frac{X_{is-x_s}}{h_s}\right)}{\sum_{i=1}^{n} \prod_{s=1}^{q_1} k\left(\frac{X_{is}-x_s}{h_s}\right) \prod_{s=q_1+1}^{q} k\left(\frac{X_{is}-x_s}{h_s}\right)} \tag{2.33}$$

如果对 $s=q_1+1,\cdots,q$ 有 $h_s=\infty$，那么对 $s=q_1+1,\cdots,q$ 有 $k((X_{is}-x_s)/h_s)=k(0)$，并且因为 $k(0)^{q-q_1}$ 从 $\hat{g}(x)$ 的分子和分母中消掉，式(2.33)简化为

$$\hat{g}(x)=\frac{\sum_{i=1}^{n} Y_i \prod_{s=1}^{q_1} k\left(\frac{X_{is}-x_s}{h_s}\right)}{\sum_{i=1}^{n} \prod_{s=1}^{q_1} k\left(\frac{X_{is}-x_s}{h_s}\right)} \tag{2.34}$$

74

因此，$\hat{g}(x)$ 将跟无关变量 $x_s(s=q_1+1,\cdots,q)$ 不相关。我们在下面的定理 2.4 证明通过交错鉴定法选择的窗宽确实具有这个性质，即对 $s=q_1+1,\cdots,q$，当 $n\to\infty$ 时，$\hat{h}_s\to\infty$。因此，交错鉴定法能够渐近地自动消除无关变量。换句话说，当某些回归元事实上无关时，交错鉴定法跟局部常数核估计量的联合能够自动降维。

交错鉴定法目标函数在现在的 $Y_i=\bar{g}(\bar{X}_i)+u_i(E(u_i|X_i)=0)$ 情况下，跟 2.2.2 节定义的是相同的，也就是条件均值函数现在仅仅依赖于相关回归元 $\bar{X}$。

类似定义于 2.2.2 节的函数 χ，现在被修正为（见习题 2.5）：

$$\bar{\chi}(a_1,\cdots,a_{q_1})=\int\left[\sum_{s=1}^{q_1} \bar{B}_s(\bar{x}) a_j^2\right]^2 \bar{f}(\bar{x}) \bar{M}(\bar{x}) \mathrm{d}\bar{x} + \frac{\kappa^{q_1}}{a_1\cdots a_{q_1}} \int \bar{\sigma}^2(\bar{x}) \bar{M}(\bar{x}) \mathrm{d}\bar{x} \tag{2.35}$$

其中

$$\bar{M}(\bar{x}) = \int \tilde{f}(x_{q_1+1},\cdots,x_q)M(\bar{x},x_{q_1+1},\cdots,x_q)\,\mathrm{d}x_{q_1+1}\cdots\mathrm{d}x_q$$

“横杠”符号表示仅仅包括 x 的前 q_1 个元素的函数。$\bar{f}(\tilde{f})$ 是 $\bar{X}(\tilde{X})$ 的边际密度，并且 $\bar{B}_s$ 除它仅仅是 $\bar{X}$ 的函数外跟 B_s 是以同样的方式定义的。注意不相关元素不再出现在 $\bar{\chi}$ 的定义中。

跟以前一样，令 $a_1^0,\cdots,a_{q_1}^0$ 表示在它们每一个都是非负的约束条件下最小化 $\bar{\chi}$ 的值。我们要求

对 $s=1,\cdots,q_1$，每一个 a_s^0 是唯一被定义的且有限的。 (2.36)

由于我们应该考虑当 $n\to\infty$ 时允许平滑参数对不相关变量从零发散，2.7 节使用的条件不能用于这里。我们使用下面对窗宽和核函数的传统约束。定义

$$H_{2n} = \left(\prod_{s=1}^{q_1} h_s\right)\prod_{s=q_1+1}^{q}\min(h_s,1)$$

令 $0<\varepsilon<1/(q+4)$。假定

75

$n^{\varepsilon-1}\leqslant H_{2n}\leqslant n^{-\varepsilon}$；$\min(h_1,\cdots,h_q)>n^{-C}$ 且对某 $C>0$ 有 $\max(h_1,\cdots,h_q)<n^C$；核 k 是一个对称、紧支撑、Hölder 连续的概率密度函数；$k(0)>k(\delta)$ 对所有 $\delta>0$ 成立。 (2.37)

如果 ε 任意小，以上关于 $h_1,\cdots,h_q$ 的条件在本质上是当 $n\to\infty$ 时，有 $nh_1,\cdots,h_{q_1}\to\infty$，$h_1,\cdots,h_{q_1}\to 0$ 和 $h_1,\cdots,h_q\to 0$，并且 h_s 趋于 0 的最快速度不能超过 n^{-a}（对某 $a>0$）并且 h_s（对无关变量）趋于 ∞ 的最快速度不超过 n^b（对某 $b>0$）。在这些条件下我们得到如下结果。

定理 2.4 在式(2.32)、式(2.36)和式(2.37)条件下，令 $\hat{h}_1,\cdots,\hat{h}_q$ 表示最小化 CV_{lc} 的平滑参数，那么

$$n^{1/(q_1+4)}\hat{h}_s \to a_s^0 \text{依概率收敛},\ 1\leqslant s\leqslant q_1$$
$$P(\hat{h}_s > C)\to 1,\ q_1+1\leqslant s\leqslant q,\ \text{所有 } C>0 \tag{2.38}$$

定理 2.4 说明对无关元素的平滑参数依概率发散到无穷，因此，所有的无关变量能够被（渐近地）自动平滑掉，而对有关变量的平滑参数继续保留了如果无关变量不存在时具有的相同的最优性质。

这里我们提到式(2.32)对定理 2.4 成立是一个充分非必要条件。Hall 等(2006)的模拟结果报告显示如果式(2.32)被条件“在 $\bar{X}$ 条件下，Y 独立于 $\tilde{X}$”代替的话，定理 2.4 的结果仍然保持成立，尽管对它的一个严格证明看起来非常有挑战性，并且仍然是未来研究的一个主题。

定理 2.4 的复杂证明由 Hall 等(2006)给出。这里我们提出一些来使得定

理 2.4 的结论成立的直觉上的理由。

我们把阶 $\eta_2 = \sum_{s=1}^{q_1} h_s^2$ 的项称作首偏误项,并且阶 $\eta_1 = (nh_1 \cdots h_q)^{-1}$ 的项称为首方差项。Hall 等(2006)证明无关变量的出现不会对 CV_{lc} 中的首偏误项产生影响,因为根据式(2.32),无关元素在比率 $\bar{\mu}_g(\bar{x}) = E[\hat{m}(x)]/E[\hat{f}(x)]$ 中消失。因此,对于 $O(\eta_2^2)$ 阶的首偏误平方项,$h_s s(q_1+1 \leqslant s \leqslant q)$ 不再出现。它们对首方差项($O(\eta_1)$ 阶项)的贡献是下面核比率的一个乘子,由于不相关元素的存在而产生(见习题 2.5):

76

$$R(\bar{x}, h_{q_1+1}, \cdots, h_q) = \frac{E\left[\left\{\prod_{s=q_1+1}^{q} k\left(\frac{\bar{x}_s - \tilde{X}_{is}}{h_s}\right)\right\}^2\right]}{\left\{E\left[\prod_{s=q_1+1}^{q} k\left(\frac{\bar{x}_s - \bar{X}_{is}}{h_s}\right)\right]\right\}^2} \tag{2.39}$$

因此,$CV_{lc}(h_1, \cdots, h_q)$ 的首项有如下形式(见习题 2.5):

$$\int \left\{\left[\sum_{s=1}^{q_1} \bar{B}_s(\bar{x}) h_s^2\right]^2 + \left[\frac{\kappa q_1 \bar{\sigma}^2(\bar{x})}{nh_1 \cdots h_{q_1} \bar{f}(\bar{x})}\right] R(\bar{x}, h_{q_1+1}, \cdots, h_q)\right\} f(x) M(x) \mathrm{d}x \tag{2.40}$$

根据 Hölder 不等式,对 $h_{q_1+1}, \cdots, h_q$ 的所有选择有 $R \geqslant 1$。很明显,当 $h_s \to \infty$ 时,$R \to 1(q_1+1 \leqslant s \leqslant q)$。事实上我们有如下的结论:

结论 2.1 当 $q_1+1 \leqslant s \leqslant q$ 时,对某 $\bar{x} \in \mathrm{supp}\omega$,使得 $R(\bar{x}, h_{q_1+1}, \cdots, h_q) = 1$ 成立的唯一值是 $h_s = \infty$。为了理解该结果,我们定义 $Z_n = \prod_{j=q_1+1}^{q} K\left(\frac{\bar{x}_s - \tilde{X}_{is}}{h_s}\right)$。那么 $\mathrm{var}(Z_n) = E[Z_n^2] - [E(Z_n)]^2 > 0$,以至于 $R = E[Z_n^2]/[E(Z_n)]^2 > 1$,除非在 Z_n 的定义中,所有的 $h_s = \infty$。在这种情况下,$Z_n \equiv k(0)^{q-q_1}$ 且 $\mathrm{var}(Z_n) = 0$,以至于当且仅当在这种情况下 $R = 1$。

因此,为最小化式(2.40)($CV_{lc}(h_1, \cdots, h_q)$ 的首项),注意平方误差和方差项都是正的,我们必有当 $n \to \infty$ 时 $R \to 1$,这意味着我们必须对发散到无穷的无关元素有平滑参数。因此,无关元素是渐近平滑掉的。在式(2.40)中使用 $R = 1$ 导出式(2.35),这就意味着对相关变量的平滑参数保留了在定理 2.4 中给出的最优性质。

77

如果使用渐近最优确定性平滑参数计算 $\hat{g}$,那么很容易证明下式依分布收敛

$$(nh_1^0 \cdots h_{q_1}^0)^{1/2} \left(\hat{g}(x) - \bar{g}(\bar{x}) - \sum_{s=1}^{q} \bar{B}_s(\bar{x})(h_s^0)^2\right) \to N(0, \bar{\sigma}_g^2(\bar{x})) \tag{2.41}$$

其中

$$B_s(\bar{x}) = \frac{\kappa_2}{2}\sum_{s=1}^{q_1}\left\{\bar{g}_{ss}(\bar{x}) + 2\frac{\bar{f}_s(\bar{x})\bar{g}_s(\bar{x})}{\bar{f}(\bar{x})}\right\}$$
$$\bar{\sigma}_g^2(\bar{x}) = \frac{\kappa^{q_1}\bar{\sigma}_u^2(\bar{x})}{\bar{f}(\bar{x})} \tag{2.42}$$

有 $\bar{\sigma}_u^2(\bar{x}) = E(u_i^2 | \bar{X}_i = \bar{x})$。

下一个定理说明当使用交错鉴定法平滑参数而不是渐近最优确定性参数时,式(2.41)给出的渐近正态性结论仍然成立。

定理 2.5 在定理 2.4 发现的同样条件下,如果 $\hat{g}(x)$ 是使用由交错鉴定法选择的平滑参数计算的,式(2.41)仍然正确,即

$$\sqrt{n\hat{h}_1\cdots\hat{h}_{q_1}}\left(\hat{g}(x) - \bar{g}(\bar{x}) - \sum_{s=1}^{q_1}\hat{h}_s^2\bar{B}_s(\bar{x})\right) \xrightarrow{d} N(0,\kappa^{q_1}\bar{\sigma}^2(\bar{x})/\bar{f}(\bar{x}))$$

定理 2.5 的证明在 Hall 等(2006)中给出。它可以被直观地理解如下:首先,注意当 $\hat{h}_s \to \infty$ $(s = q_1 + 1,\cdots,q)$ 时,我们有 $\hat{g}(x,\hat{h}_1,\cdots,\hat{h}_q) = \hat{g}(\bar{x},\hat{h}_1,\cdots,\hat{h}_{q_1}) + (\text{s. o.})$,其中 $\hat{g}(\bar{x},\hat{h}_1,\cdots,\hat{h}_{q_1})$ 是 $\bar{g}(\bar{x})$ 仅仅使用相关回归元的一个核估计量。其次,给定事实 $\hat{h}_s - h_s^0 = o_p(h_s^0)$ $(s = 1,\cdots,q_1)$,我们期望 $\hat{g}(x,\hat{h}_1,\cdots,\hat{h}_q) = \hat{g}(\bar{x},h_1^0,\cdots,h_{q_1}^0) + o_p\left(\sum_{s=1}^{q_1}(h_s^0)^2\right)$。定理 2.5 来自这个结果和式(2.41)。

注意条件(2.32)是非常强的。它不仅假定 $\tilde{X}$ 独立于 Y,也要求 $\tilde{X}$ 是独立于 $\bar{X}$ 的。理想地,人们愿意放松这个条件为(i) $E[Y|\bar{X},\tilde{X}] = E[Y|\bar{X}]$ 几乎确定成立,或者(ii)在条件 $\bar{X}$ 下,Y 是独立于 $\tilde{X}$ 的。Hall 等(2006)猜测当式(2.32)被放松并且被以上条件(i)或(ii)取代时,定理 2.4 仍然成立。Hall 等报告的模拟结果中提供了一些证据来支持这个猜测。

2.2.5 关于交错鉴定法的一些进一步结论

78

Racine 和 Li(2004)也证明了 $\hat{h}_s/h_s^0 - 1$ 的收敛速度,它由下式给出

$$\frac{\hat{h}_s - h_s^0}{h_s^0} \equiv \frac{\hat{h}_s}{h_s^0} - 1 = O_p(n^{-\alpha/(4+q)}) \tag{2.43}$$

其中 $\alpha = \min\{q/2,2\}$。当 $q = 1$ 时,我们得到 $(\hat{h}_s - h_s^0)/h_s^0 = O_p(n^{-1/10})$,这跟 Härdle 等(1988)得到的结果是一致的。式(2.43)意味着 $\hat{h}$ 依概率收敛于非随机的最优平滑参数 h_s^0。

式(2.43)表明,$(\hat{h}_s - h_s^0)/h_s^0 = O_p(n^{-q/[2(4+q)]})$ 对 $q \leqslant 4$ 成立,以及

$(\hat{h}_s - h_s^0)/h_s^0 = O_p(n^{-2/(4+q)})$对 $q \geqslant 4$ 成立。对$(\hat{h}_s - h_s^0)/h_s^0$ 有两个不同表达式的原因(取决于 $q \geqslant 4$ 是否成立)是,在交错鉴定函数 $CV_{lc}(h_1, \cdots, h_q)$的高阶展开式,我们有阶 $O_p\left(\sum_{s=1}^{q} h_s^6\right)$ 项和一个如下形式的二阶退化 U 统计量

$$(n(n-1))^{-1} \sum_i \sum_{j \neq i} v_i v_j K_h(X_i, X_j)$$

其中 $E(v_i | X_i) = 0$,它有阶 $O_p(n(h_1 \cdots h_q)^{1/2})^{-1}$;见附录 A 对退化 U 统计量的一个一般处理。当 $q \leqslant 4$ 时,$O_p(n(h_1 \cdots h_q)^{1/2})^{-1}$项占优于 $O_p(h_s^6)$,因为 $h_s \sim O_p(h_s^0) = O_p(n^{-1/(4+q)})$,而当 $q \geqslant 5$ 时,$O_p(h_s^6)$变为占优项。因此,收敛速度不同且依赖于 $q \leqslant 4$ 还是 $q \geqslant 5$(见 Racine 和 Li(2004)的详细证明)。

2.3 一致收敛速度

沿着在第 1 章 1.12 节讨论的关于推导密度函数估计量的几乎确定一致收敛速度时使用的那些类似的思路,我们能够建立 $\hat{g}(x)$对 $g(x)$的几乎确定一致收敛速度($x \in \mathcal{S}$),其中$\mathcal{S}$是一个$\mathbb{R}^q$ 空间的紧集,它排除了 X 支撑的边界范围。

条件 2.1 假定

79

(i) $f(x)$是可微的,而且 $g(x)$是二阶可微的,并且导函数都满足 Lipschitz 条件$|m(x) - m(v)| \leqslant C|x - v|$,对某 $C > 0$($m(\cdot) = g_{ss}(\cdot)$或$f(\cdot)$)。

(ii) $\sigma^2(x) = E(u_i^2 | x)$是一个连续函数,且 $\inf_{x \in \mathcal{S}} f(x) \geqslant \delta > 0$。

(iii) 核 $k(\cdot)$是对称、有界且有紧支撑(即 $k(v) = 0$ 对 $|v| > 1$)的。定义 $H_l(v) = |v|^l K(v)$,我们假定$|H_l(v) - H_l(u)| \leqslant C_2 |u - v|$,对所有 $0 \leqslant l \leqslant 3$。

定理 2.6 在条件 2.1 下,我们有

$$\sup_{x \in \mathcal{S}} |\hat{g}(x) - g(x)| = O\left(\frac{(\ln n)^{1/2}}{(nh_1 \cdots h_q)^{1/2}} + \sum_{s=1}^{q} h_s^2\right) \text{ a.s.} \tag{2.44}$$

式(2.44)的证明跟第 1 章定理 1.4 的证明非常类似。通过记 $\hat{g}(x) - g(x) = \hat{m}(x)/\hat{f}(x)$,其中 $\hat{m}(x) = (\hat{g}(x) - g(x))\hat{f}(x)$,能够证明

$$\sup_{x \in \mathcal{S}} |\hat{m}(x)| = O\left(\sum_{s=1}^{q} h_s^2 + \frac{(\ln n)^{1/2}}{(nh_1 \cdots h_q)^{1/2}}\right) \text{ a.s.}$$

同样,在$\mathcal{S}$中密度不为零时,这个事实以及定理 1.4 意味着几乎处处 $\inf_{x \in \mathcal{S}} \hat{f}(x) \geqslant \delta' > 0$,其中 δ' 是一个正常数。因此,

$$\begin{aligned}\sup_{x \in \mathcal{S}} |\hat{g}(x) - g(x)| &= \sup_{x \in \mathcal{S}} |\hat{m}(x)/\hat{f}(x)| \leqslant \sup_{x \in \mathcal{S}} |\hat{m}(x)| / |\inf_{x \in D} \hat{f}(x)| \\ &= O\left((\ln n)^{1/2}/(nh_1 \cdots h_q)^{1/2} + \sum_{s=1}^{q} h_s^2\right) \text{ a.s.}\end{aligned}$$

在细节上考察了局部常数核估计量后,我们现在转向另一种流行的方法——局部多项式方法。

2.4 局部线性核估计

尽管局部常数估计量是核回归的经典方法,但它不是没有缺陷的。尤其是当在支撑边界附近估计一个回归函数时它有潜在较大的偏差。另一方面,局部线性估计量由 Stone 和 Cleveland(1979)提出,尽管它具有局部常数方法的许多性质(如它们的方差是相同的),然而它是修正边界问题最为人熟知的方法之一,因为它的偏差不是指定密度 $f(x)$ 的一个函数(见 Fan(1992), Fan(1993),以及 Fan 和 Gijbels(1992))。对局部线性估计量的一个详尽的处理参见 Fan 和 Gijbels(1996)的优秀专著。 80

我们再一次考虑如下形式的一个回归模型

$$Y_j = g(X_j) + u_j, \quad j = 1, \cdots, n \tag{2.45}$$

回忆在 2.1 节讨论的局部常数核估计量是由下式给出的

$$\hat{g}(x) = \frac{\sum_{j=1}^{n} Y_j K\left(\frac{X_j - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{X_j - x}{h}\right)} \tag{2.46}$$

它也可以作为如下最小化问题的解 a 来得到:

$$\min_{a} \sum_{j=1}^{n} (Y_j - a)^2 K\left(\frac{X_j - x}{h}\right) \tag{2.47}$$

令 $\hat{a} = \hat{a}(x)$ 是最小化式(2.47)的解,人们能容易看到 $\hat{a} \equiv \hat{g}(x)$,其定义见式(2.46)。

然而,注意式(2.47)使用一个常数 a 在 x 的邻域逼近 $g(x)$(或 Y),因为式(2.47)仅仅使用 Y_j 的一个局部平均(接近 x 的 X_j)来估计 $g(x)$。由于这个原因,定义于式(2.46)的 $\hat{g}(x)$ 被称为一个"局部常数"核估计量。然而,人们可以另外使用一个"局部线性"(或高阶多项式)估计量来估计 $g(x)$。局部线性方法的一个特征是它为 $g(x)$ 的一阶导数 $g^{(1)}(x) \stackrel{\text{def}}{=} \partial g(x)/\partial x$ 自动提供了一个估计量,尽管局部常数的偏导数估计量得到了很好的研究;进一步细节可参见 Vinod 和 Ullah(1988), Rilstone 和 Ullah(1989), Härdle 和 Stoker(1989),以及 Pagan 和 Ullah(1999,4.2 节)。

局部线性方法是基于如下的最小化问题:

$$\min_{\{a,b\}} \sum_{j=1}^{n} (Y_j - a - (X_j - x)'b)^2 K\left(\frac{X_j - x}{h}\right) \tag{2.48}$$

81 令 $\hat{a}=\hat{a}(x)$ 和 $\hat{b}=\hat{b}(x)$ 是式(2.48)的解。我们在这一节将证明 $\hat{a}(x)$ 是 $g(x)$ 的一个一致估计量，且 $\hat{b}(x)$ 是 $g^{(1)}(x)=\partial g(x)/\partial x$ 的一个一致估计量(b 和 $g^{(1)}(x)$ 都是 $q\times 1$ 向量)。式(2.48)给出的局部线性方法是容易理解的，因为它类似于“局部最小二乘”估计量。因此斜率估计量 $\hat{b}$ 估计了局部斜率 $g^{(1)}(x)$。

令 $\delta=\delta(x)=(a(x),(b(x))')'$，令 $\mathcal{Y}$ 是有第 i 个元素为 Y_i 的 $n\times 1$ 维向量，令 $\mathcal{X}$ 是一个第 i 行为 $(i,(X_i-x)')$ 的 $n\times(1+q)$ 矩阵，并且令 $\mathcal{K}(x)$ 是一个第 i 个对角元素为 $K\left(\frac{X_i-x}{h}\right)$ 的 $n\times n$ 对角矩阵。那么用向量-矩阵符号，式(2.48)可以被写为

$$\min_{\{a,b\}}(\mathcal{Y}-\mathcal{X}\delta)'\mathcal{K}(x)(\mathcal{Y}-\mathcal{X}\delta) \tag{2.49}$$

方程式(2.49)是一个标准的广义最小二乘问题。令 $\hat{\delta}=(\hat{a},\hat{b}')'$ 是(2.49)的解。那么对一个广义最小二乘(GLS)估计量使用标准公式，我们得到

$$\begin{aligned}\hat{\delta}(x) &= (\mathcal{X}'\mathcal{K}(x)\mathcal{X})^{-1}\mathcal{X}'\mathcal{K}(x)\mathcal{Y}\\ &= \left[\sum_{j=1}^{n}K\left(\frac{X_j-x}{h}\right)\begin{pmatrix}1\\ X_j-x\end{pmatrix}(1,(X_j-x)')\right]^{-1}\\ &\quad\times\sum_{j=1}^{n}K\left(\frac{X_j-x}{h}\right)\begin{pmatrix}1\\ X_j-x\end{pmatrix}Y_j\end{aligned} \tag{2.50}$$

如下条件对在式(2.50)中给出的 $\hat{\delta}(x)$ 建立一致性和渐近正态性是必需的。

条件 2.2

(i) $\{X_j,Y_j\}_{j=1}^{n}$ 是独立同分布，$g(x)$ 和 $f(x)$ 以及 $\sigma^2(x)=E(u_i^2|x)$ 都是二阶可微的。

(ii) K 是一个有界二阶核。

(iii) 当 $n\to\infty$ 时，$nh_1\cdots h_q\sum_{s=1}^{q}h_s^2\to\infty$ 且 $nh_1\cdots h_q\sum_{s=1}^{q}h_s^6\to 0$。

下一个定理建立了 $\hat{\delta}(x)$ 的渐近正态性。

82 **定理 2.7** 回忆 $\kappa=\int k(v)^2\mathrm{d}v$，$\kappa_2=\int k(v)v^2\mathrm{d}v$，定义 $\kappa_{22}=\int v^2k(v)^2\mathrm{d}v$，并且令

$$D(n)=\begin{bmatrix}\sqrt{nh_1\cdots h_q}, & 0\\ 0, & \sqrt{nh_1\cdots h_q}D_h\end{bmatrix}$$

$$\sum=\begin{bmatrix}\kappa^q\sigma^2(x)/f(x), & 0\\ 0, & \kappa^{q-1}\kappa_{22}\sigma^2(x)I_q/[\kappa_2^2f(x)]\end{bmatrix}$$

其中 D_h 是第 s 个对角元素由 h_s 给出的一个 $q \times q$ 对角矩阵，并且其中 I_q 是一个 q 维的单位矩阵。那么在条件 2.2 下我们有

$$D(n)\left(\hat{\delta}(x)-\delta(x)-\begin{bmatrix}\frac{\kappa_2}{2}\sum_{s=1}^{q}h_s^2 g_{ss}(x)\\0\end{bmatrix}\right)\xrightarrow{d}N\left(0,\sum\right) \tag{2.51}$$

定理 2.7 的证明在 2.7.2 节给出。

注意定理 2.7 意味着如下不同的收敛速度：

$$(nh_1\cdots h_q)^{1/2}\left[\hat{a}(x)-g(x)-\frac{\kappa_2}{2}\sum_{s=1}^{q}h_s^2 g_{ss}(x)\right]\xrightarrow{d}N\left(0,\frac{\kappa^q\sigma^2(x)}{f(x)}\right) \tag{2.52}$$

且

$$(nh_1\cdots h_q)^{1/2}h_s[\hat{b}_s(x)-g_s(x)]\xrightarrow{d}N\left(0,\frac{\kappa^{q-1}\kappa_{22}\sigma^2(x)}{\kappa_2^2 f(x)}\right),s=1,\cdots,q \tag{2.53}$$

其中 $g_s(x)=\partial g(x)/\partial x_s$ 是 $\partial g(x)/\partial x$ 的第 s 个元素。

令 $\hat{g}(x)=\hat{a}(x)$ 且 $\hat{g}_s(x)=\hat{b}_s(x)$ 分别表示 $g(x)$ 和 $g_s(x)$ 的局部线性核估计。那么 $\hat{g}(x)-g(x)=O_p(\eta_2+\eta_1^{1/2})$，并且 $\hat{g}_s(x)-g_s(x)=O_p(\eta_2+\eta_1^{1/2}h_s^{-1})$ $(\eta_2=\sum_{s=1}^{q}h_s^2,\eta_1=(nh_1\cdots h_q)^{-1})$。这是一个标准结果，即导数估计的收敛速度慢于回归函数估计的收敛速度。为估计 $g(x)$ 的第 l 阶导数，收敛速度将是 $O_p\left(\eta_2+\eta_1^{1/2}\left(\sum_{s=1}^{q}h_s^{2l}\right)^{-1/2}\right)$。我们在 2.5 节讨论高阶导数估计量。

为了使偏误和方差都收敛于 0，$g(x)$ 的一致估计要求 $h_s\to 0(s=1,\cdots,q)$ 并且 $nh_1\cdots h_q\to\infty$；而为了使导数估计量的方差项收敛于 0，$g_s(x)$ 的一致估计要求一个更强的条件，也就是 $nh_1\cdots h_q\sum_{s=1}^{q}h_s^2\to\infty$。

注意当潜在回归函数事实上对 x 是线性时（即 $g(x)=\alpha_0+x'\alpha_1$），其中标量 α_0 和 $q\times 1$ 向量 α_1 是常数），局部线性估计量对 $(h_1,\cdots,h_q)$ 的任意值表现出偏误项为 0 的性质（见习题 2.10）。因此，当 $g(x)$ 对 x 线性时，局部线性估计量是无偏的。当偏误为 0 时，我们能够考虑 $h_s=\infty$ $(s=1,\cdots,q)$，并且很容易证明在这种情况下局部线性估计量退化为 $\hat{a}(x)=\hat{\alpha}_0+x'\hat{\alpha}_1$，其中 $\hat{\alpha}_0$ 和 $\hat{\alpha}_1$ 分别是 α_0 和 α_1 的普通最小二乘估计量（见习题 2.6）。因此，如果人们允许对所有的 $s=1,\cdots,q$，对 h_s 可以取充分大的值，局部线性估计量把最小二乘估计量视为一种特殊情况。在下一节当我们讨论为局部线性估计量的交错鉴定选择平滑参

83

数时,我们再讨论这一点,它对窗宽选择和收敛速度都有非常重要的意义。

2.4.1 局部线性窗宽选择:最小二乘交错鉴定法

令 $\hat{g}_{-i,L}(X_i)$ 表示去一局部线性估计量。也就是令 $(\hat{a}_i,\hat{b}_i)$ 在如下最小化问题中 (a,b) 的解:

$$\min_{\{a,b\}}\sum_{j\neq i,j=1}^{n}[Y_j-a-(X_j-X_i)'b]^2K\left(\frac{X_i-X_j}{h}\right)$$

其中 $K\left[\frac{X_i-X_j}{h}\right]=\prod_{s=1}^{q}k\left(\frac{X_{is}-X_{js}}{h_s}\right)$。那么 $\hat{a}_i\equiv\hat{g}_{-i,L}(X_i)$ 是 $g(X_i)$ 的去一局部线性核估计量。

窗宽选择的局部线性交错鉴定法是选择最小化下式的那些 h_s

$$\mathrm{CV}_{ll}(h_1,\cdots,h_q)=\min_h\frac{1}{n}\sum_i[Y_i-\hat{g}_{-i,L}(X_i)]^2M(X_i)\tag{2.54}$$

其中 $M(\cdot)$ 是一个权函数。注意目标函数(2.54)仅仅包括 $g(\cdot)$ 的估计而没有它的导数,因此我们仅仅需要在式(2.74)中提出的关于 $h_1,\cdots,h_q$ 的条件,也就是说我们不需要关于平滑参数的更强条件以保证 $g(x)$ 导数的一致估计。

84

对固定的 $x\in\mathbb{R}^q$,我们已经推导了局部线性估计量 $\hat{g}(x)=\hat{a}(x)$ 的渐近偏误和方差项,而式(2.52)意味着

$$E[\hat{g}(x)-g(x)]^2=\left[\frac{\kappa_2}{2}\sum_{s=1}^{q}g_{ss}(x)h_s^2\right]^2+\frac{\kappa^q}{nh_1\cdots h_q}\frac{\sigma^2(x)}{f(x)}+o(\eta_2^2+\eta_1)\tag{2.55}$$

如我们在 2.2.2 节讨论的,交错鉴定目标函数渐近等价于 $\int E[\hat{g}(x)-g(x)]^2f(x)M(x)\mathrm{d}x$。使用式(2.55),人们可能猜测到 CV_{ll} 的首项将是

$$\begin{aligned}\mathrm{CV}_{ll,0}&\sim\int E[\hat{g}(x)-g(x)]^2f(x)M(x)\mathrm{d}x\\&=\int\left[\frac{\kappa_2}{2}\sum_{s=1}^{q}g_{ss}(x)h_s^2\right]^2f(x)M(x)\mathrm{d}x+\frac{\kappa^q\int\sigma^2(x)M(x)\mathrm{d}x}{nh_1\cdots h_q}\\&\quad+o(\eta_2^2+\eta_1)\end{aligned}\tag{2.56}$$

并且当 CV_{ll} 的首项确实由式(2.56)给出时,这个猜测的结果是正确的(见 Li 和 Racine(2004a)的一个详细证明)。

类似地,使用 2.2.2 节提出的方法,我们把式(2.56)的首项表示为 $n^{-4/(q+4)}\chi_{ll}\cdot(a_1,\cdots,a_q)$,其中 a_s 被定义为 $h_s=a_sn^{-1/(q+4)}(s=1,\cdots,q)$,并且

$$\chi_{ll}(a_1,\cdots,a_q)=\int\left[\frac{\kappa_2}{2}\sum_{s=1}^{q}g_{ss}(x)a_s^2\right]^2f(x)M(x)\mathrm{d}x$$

$$+\frac{\kappa^q}{a_1\cdots a_q}\int\sigma^2(x)M(x)\mathrm{d}x \tag{2.57}$$

令 $a^0_{1,l},\cdots,a^0_{q,l}$表示最小化$\chi_{ll}$的那些 $a_1,\cdots,a_q$ 的值并且假定

每一个 $a^0_{s,l}$是唯一被定义的且是正的和有限的。 (2.58)

假定(2.58)排除了 $a^0_{s,l}=\infty$ 的情况,这意味着 $g_{ss}(x)$不能是一个零函数。也就是说我们明确排除了 $g(x)$在它的任何一个元素 x_s 是线性的情况。

令 $\hat{h}_s$ 表示最小化式(2.54)的 h_s 的值,Li 和 Racine(2004a)证明

$$n^{1/(q+4)}\hat{h}_s\to a^0_{s,l}\text{依概率收敛},\ s=1,\cdots,q \tag{2.59}$$

85

式(2.59)说明局部线性交错鉴定平滑参数收敛于最优平滑参数,并且相应局部线性估计量的收敛速度跟当潜在回归函数在它的任意一个元素不是线性时的局部常数交错鉴定情况相同。

当 $g(x)$在 x 的某些元素是线性时,比如 x_s,那么假定(2.58)不再成立。在这种情况下,跟在 2.2.4 节一样,我们也需要修正施加于平滑参数的条件,以至于我们允许对应于线性进入 $g(x)$ 的那些回归元的平滑参数发散到无穷。那么,我们将期望当 $g(x)$在 x 是线性时,局部线性最小二乘交错鉴定应该有能力选择一个较大的 h_s 值,而当回归元非线性进入时选择相对较小的 h_t 值。Li 和 Racine(2004a)报告的模拟结果证明确实是这种情况,尽管我们没有看到对这个结果提供一个严格理论证明的任何工作。

具体地,让我们考虑 $q=2$ 且潜在回归函数 $g(x_1,x_2)=\theta(x_1)+x_2\alpha$ 是一个局部线性函数(例如,关于 x_2 线性)的情况。在这种情况下,局部线性交错鉴定法将倾向于选择一个非常大的 $\hat{h}_2$($\hat{h}_2\to\infty$)值,而 $\hat{h}_1=O_p(n^{-1/5})$(不是 $O_p(n^{-1/6})$)。因此局部线性交错鉴定法将建议 $g(\cdot)$关于 x_2 是线性的。

2.5 局部多项式回归

2.5.1 单变量情况

正如当我们推导局部线性核估计量时的情况,我们能使用一个更高阶的局部多项式估计量来估计 $g(x)$。当 x 是一个单变量,一个 p 阶局部多项式核估计量是基于最小化如下的目标函数:

$$\min_{\{b_0,b_1,\cdots,b_p\}}\sum_{j=1}^{n}(Y_j-b_0-b_1(X_j-x)-\cdots-b_p(X_j-x)^p)^2\times K\left(\frac{X_j-x}{h}\right) \tag{2.60}$$

86 令 $\hat{b}_l$ 代表最小化式(2.60)的 $b_l(l=0,1,\cdots,p)$ 值。那么 $\hat{b}_0$ 估计 $g(x)$，且 $l!\,\hat{b}_l$ 估计 $g^{(l)}(x)$，其中 $g^{(l)}(x)=\mathrm{d}^l g(x)/\mathrm{d}x^l$ 是 $g(x)$ 的 $l(l=1,\cdots,p)$ 阶导数。

再一次，式(2.60)是一个标准的加权最小二乘问题。令 $\mathcal{X}_j=(1,(X_j-x),\cdots,(X_j-x)^p)'$，那么

$$\hat{b}=\begin{bmatrix}\hat{b}_0\\ \hat{b}_1\\ \vdots\\ \hat{b}_p\end{bmatrix}=\left[\sum_j \mathcal{X}_j\mathcal{X}_jK\left(\frac{X_j-x}{h}\right)\right]^{-1}\sum_j \mathcal{X}_jY_jK\left(\frac{X_j-x}{h}\right) \tag{2.61}$$

Ruppert 和 Wand(1994)研究了 $\hat{b}_0$ 的主要条件偏误和条件方差，我们把它总结在下面的定理中。

定理 2.8 令 $\hat{g}(x)\equiv\hat{b}_0(x)$，且令 $Z_n=\{X_i\}_{i=1}^n$。

如果 p 为奇数，

$$E[\hat{g}(x)\mid\mathcal{Z}_n]-g(x)=h^{p+1}c_{1,p}\left[\frac{m^{(p+1)}(x)}{(p+1)!}\right]+o(h^{p+1}) \tag{2.62}$$

而如果 p 为偶数，

$$E[\hat{g}(x)\mid\mathcal{Z}_n]-g(x)=h^{p+2}\left\{c_{2,p}\left[\frac{m^{(p+1)}(x)f^{(1)}(x)}{f(x)(p+1)!}\right]+c_{3,p}\left[\frac{m^{(p+2)}(x)}{(p+2)!}\right]\right\}+o(h^{p+2}) \tag{2.63}$$

在任意一种情况下，

$$\operatorname{var}(\hat{g}(x)\mid\mathcal{Z}_n)=\left[\frac{c_{4,p}\sigma^2(x)}{nhf(x)}\right]+o\left(\frac{1}{nh}\right) \tag{2.64}$$

其中 $c_{j,p}(j=1,2,3,4)$ 是 Ruppert 和 Wand(1994)中定义的一些常数。

定理 2.8 说明主要条件偏误项依赖于 p 为奇数还是偶数。通过一个 Taylor 级数展开论证，我们知道当考虑 $|X_j-x|\leqslant h$ 时，一个 p 阶多项式展开的余项应该是 $O(h^{p+1})$ 阶，因此对奇数 p 的结果是很容易理解的。当 p 为偶数时，$p+1$

87 是奇数，因此 h_{p+1} 项跟 l 为奇数时 $\int K(v)v^l\mathrm{d}v$ 是相关的，并且因为 $K(v)$ 是一个偶函数，该项为 0。因此，h^{p+1} 项消失，而余项变为 $O(h^{p+2})$。由于 p 要么为奇数要么为偶数，我们看到偏误项是 h 的一个偶数幂。这跟人们使用高阶核函数的局部常数估计量情况类似，它基于一个对称核函数(一个偶函数)，其中偏误总是 h 的一个偶数幂。

概括起来，在条件 $\mathcal{X}_n$ 下，我们有

$$\begin{aligned}&\text{当}\, p\, \text{为奇数时}, \hat{g}(x)-g(x)=O_p(h^{p+1}+(nh)^{-1/2})\\&\text{当}\, p\, \text{为偶数时}, \hat{g}(x)-g(x)=O_p(h^{p+2}+(nh)^{-1/2})\end{aligned} \tag{2.65}$$

令 $\hat{g}^{(l)}(x)=l!\hat{b}_l(x)$ 表示基于一个 p 阶局部多项式拟合($l\leqslant p$)的 $g^{(l)}(x)$ 估计量。下面的定理说明了 $g^{(l)}(x)$ 的主要偏误和方差项。

定理 2.9 当 $p-l$ 为奇数时,

$$E[\hat{g}^{(l)}(x)\mid \mathcal{X}_n]-g^{(l)}(x)=h^{p-l+1}c_{1,l,p}\left[\frac{m^{(p+1)}(x)}{(p+1)!}\right]+o(h^{p-l+1})$$

当 $p-l$ 为偶数时,

$$\begin{aligned}E[\hat{g}^{(l)}(x)\mid \mathcal{X}_n]-g^{(l)}(x)=h^{p-l+2}\Bigg\{&c_{2,l,p}\left[\frac{m^{(p+2)}(x)}{(p+2)!}\right]\\&\times\int v^{p+2}K_{(l,p)}(v)\,\mathrm{d}v\\&+c_{3,l,p}\left[\frac{m^{(p+1)}(x)f^{(1)}(x)}{f(x)(p+1)!}\right]\Bigg\}\end{aligned}$$

在任意一种情况下,

$$\operatorname{var}[\hat{g}^{(l)}(x)\mid \mathcal{X}_n]=\left[\frac{c_{4,l,p}\sigma^2(x)}{nh^{2l+1}f(x)}\right]+o((nh^{2l+1})^{-1})$$

其中 $c_{j,l,p}(j=1,2,3,4)$ 是在 Ruppert 和 Wand(1994)中定义的常数。

注意如果 $p-l$ 是奇数,$f^{(1)}(x)/f(x)$ 没有出现在条件偏误项中。Ruppert 和 Wand(1994)也证明当 $p-l$ 是奇数时,在边界处的偏误跟位于内部的点的偏误有相同的阶。因此,当 $p-l$ 是奇数时,局部多项式均值估计有吸引力的边界行为扩展到导数估计。然而,当 $p-l$ 是偶数时,在边界处的偏误比在内部的点大,且偏误在 $f(x)$ 不连续的点也变大。出于这些原因,当估计 $g^{(l)}(x)$ 时,我们建议严格设定 $p-l$ 为奇数。

88

定理 2.9 意味着,在条件 $\mathcal{X}_n$ 下,

$$\begin{aligned}&\text{当}\, p\, \text{为奇数时}, \hat{g}^{(l)}(x)-g^{(l)}(x)=O_p(h^{p-l+1}+(nh^{2l+1})^{-1/2})\\&\text{当}\, p\, \text{为偶数时}, \hat{g}^{(l)}(x)-g^{(l)}(x)=O_p(h^{p-l+2}+(nh^{2l+1})^{-1/2})\end{aligned} \tag{2.66}$$

使用 Liapunov 中心极限定理,我们也能得到 $\hat{g}(x)$ 和 $\hat{g}^{(l)}(x)$ 的渐近正态性。我们把对渐近正态性结果的讨论推迟到下一节,在那里我们将讨论多元局部多项式回归。

2.5.2 多变量情况

对多变量回归元情况下的一个 p 阶局部多项式估计量是更不方便用符号来说明的。Masry(1996a,1996b)研究了多变量情况,为 $g(x)$ 的局部多项式估计

量及其直到 p 阶的导数提出了一些经过仔细考虑的符号并得到了几乎确定一致的收敛速度和逐点渐近正态性结果。借鉴 Masry(1996b),我们引入如下的符号:

$$r=(r_1,\cdots,r_q),\quad r!=r_1!\times\cdots\times r_q!,\quad \bar{r}=\sum_{j=1}^{q}r_j \tag{2.67}$$

$$x^r=x_1^{r_1}\times\cdots\times x_q^{r_q},\quad \sum_{0\leqslant\bar{r}\leqslant p}=\sum_{j=0}^{p}\sum_{r_1=0}^{j}\cdots\sum_{r_q=0}^{j}$$
$$(\bar{r}\equiv r_1+\cdots+r_q=j) \tag{2.68}$$

和

$$(D^r g)(x)=\frac{\partial^r g(x)}{\partial x_1^{r_1}\cdots\partial x_q^{r_q}} \tag{2.69}$$

使用这种符号,并假定 $g(x)$ 在点 x 存在直至 $p+1$ 阶导数,我们能够使用一个由下式给出的阶为 p 的多元多项式局部逼近 $g(z)$

89

$$g(z)\cong\sum_{0\leqslant\bar{r}\leqslant p}\frac{1}{r!}(D^r)g(v)\big|_{v=x}(z-x)^r \tag{2.70}$$

定义一个多元加权最小二乘函数:

$$\sum_{i=1}^{n}\Big\{Y_i-\sum_{0\leqslant\bar{r}\leqslant p}b_r(x)(X_i-x)^r\Big\}^2K\Big(\frac{X_i-x}{h}\Big) \tag{2.71}$$

关于每一个 b_r 最小化式(2.71)给出一个 $\hat{b}_r(x)$ 的估计,并且根据式(2.70),我们知道 $r!\hat{b}_r(x)$ 估计 $(D^r g)(x)$ 以至于 $(D^r\hat{g})(x)=r!\hat{b}_r(x)$。对局部线性核估计量的多变量窗宽选择的一个一般处理,参见 Yang 和 Tschernig (1999)。

2.5.3 局部多项式估计量的渐近正态性

局部 p 阶多项式回归可以用来估计出 $g(x)$ 直到 p 阶的导数。我们用 N_l 来表示 $g(x)$ 不同的 l 阶导数的个数。例如,$N_0=1$,$N_1=q$(q 个不同的一阶导数),并且 $N_2=q(q+1)/2$(q 个二阶自导数和 $q(q-1)/2$ 个交叉二阶导数)。一般公式是

$$N_l=\begin{bmatrix}q+l-1\\q-1\end{bmatrix}=\frac{(q+l-1)!}{(q-1)!l!},\quad l=0,1,\cdots,p$$

$g(x)$ 有 N_l 个不同的 l 阶导数,并且我们使用 $\nabla^{(l)}g(x)$ 来表示这些 l 阶导数的 $N_l\times 1$ 向量,使用字典式排序。例如,当 $l=2$ 时,我们有 $N_2=q(q+1)/2$ 个不同的二阶导数,且 $\nabla^{(2)}g(x)$ 由下式定义

$$(\nabla^{(2)} g)(x) = \begin{bmatrix} \frac{\partial^2 g(x)}{\partial x_1^2} \\ \frac{\partial^2 g(x)}{\partial x_1 \partial x_2} \\ \vdots \\ \frac{\partial^2 g(x)}{\partial x_q^2} \end{bmatrix}$$

Masry(1996b)使用如下的条件来得到局部多项式估计量的渐近正态性结果。

条件 2.3

90

(i) $g(x)$有直到 $p+1$ 阶的连续导数。

(ii) $k(v)$是一个有紧支撑的有界二阶核函数。

(iii) 为了表达方便,令 $h_1 = \cdots = h_q = h$,假定 $h = O(n^{-1/(q+2p+2)})$。

定理 2.10 在条件 2.3 下,对 $0 \leqslant l \leqslant q$ 我们有

$$(nh^{q+2l})^{1/2}[\nabla^{(l)}\hat{g}(x) - \nabla^{(l)}g(x) - A_{l,p+1}(x)h^{p+1-l}] \xrightarrow{d} N\left(0, \frac{\sigma^2(x)}{f(x)}V_l\right)$$

在 $\sigma^2(x)$和$f(x)$的连续点有$f(x) > 0$,其中 $m_{p+1}(x)$、A_l 和 V_l 的定义在 2.7 节给出。

证明:见 Masry(1996a)的定理 5。 □

如果我们把定理 2.10 跟定理 2.8 和 2.9 比较,我们注意到定理 2.10 没有产生取决于 $p-l$ 是奇数还是偶数的不同偏误表达式。能够证明当 $q=1$ 时(单回归元情况),定理 2.10 的偏误表达式跟定理 2.8 和 2.9 当 $p-l$ 是奇数的偏误表达一致。当 $p-l$ 是偶数时,定理 2.10 给出的首偏误项实际上为 0。因此,当 $p-l$ 是偶数时,非零首偏误项将有阶 $O(h^{p+2-l})$。在 2.7 节我们对 $p=1$ 的情况证明了这点,那里我们将证明 $A_{l,p+1}(x) = A_{1,2}(x) = 0$,所以,未消失的首偏误项将有阶 $O(h^{p+2-l}) = O(h^2)$,而不是 $O(h^{p+1-l}) = O(h)$ $(p = l = 1)$。这跟定理 2.7 的结果一致,因此在定理 2.10 给出的结果跟在定理 2.8 和 2.9 中给出的结论是一致的。

假定 $p-l$ 是奇数,我们有

$$\text{MSE}(\hat{g}^{(l)}(x)) = O(h^{2(p-l+1)} + (nh^{q+2l})^{-1})$$

或者

$$\hat{g}^{(l)}(x) = O_p(h^{p-l+1} + (nh^{q+2l})^{-1/2})$$

Masry(1996b)得到了局部多项式估计量的几乎确定一致收敛速度,我们在下一个定理中报告他的结果。

定理 2.11 对每一个 $0 \leqslant l \leqslant p$,我们几乎确定有

91

$$\sup_{x\in S}|\nabla^{(l)}\hat{g}(x)-\nabla^{(l)}g(x)|=O\left(\left(\frac{\ln(n)}{nh^{q}+2l}\right)^{1/2}+h^{p-l+1}\right) \tag{2.72}$$

证明:见 Masry(1996b)的定理 6。 □

跟我们以前对这个问题的讨论一样,以上的结果说明,当 $p-l$ 是偶数时,偏误的阶不是尽可能地小。例如,当 $p-l$ 是偶数时,定理 2.11 提供了一个阶为 $O(h^{p-l+1})$ 的偏误。当 $p-l$ 是偶数时,给定 $O(h^{p-l+2})$ 阶的逐点速度,应该把偏误项从 $O(h^{p-l+1})$ 减小为 $O(h^{p-l+2})$,尽管对于给定模型的一般结构这可能相当复杂。

基于(最小最大)效率的争论,Ruppert 和 Wand(1994)建议 $p-l$ 应该被选择为奇数。当 $p-l$ 是奇数时,定理 2.11 偏误的阶跟逐点速度结果的偏误的阶一致。

对 $l=0$,定理 2.11 给出一致速度为

$$\begin{aligned}|\nabla^{(0)}\hat{g}(x)-\nabla^{(0)}g(x)|&\equiv|\hat{g}(x)-g(x)|\\&=O((\ln(n)/(nh^{q}))^{1/2}+h^{p+1})\end{aligned}$$

其中 $\hat{g}(x)$ 是 $g(x)$ 的 p 阶局部多项式估计量。我们看到在 p 阶局部多项式回归中的偏误是 $O(h^{p+1})$ 阶,并且当 p 为偶数时,它可被进一步减为 $O(h^{p+2})$。对 $p\geq 2$,偏误项类似于对局部常数估计量使用一个更高阶核函数的情况。

对 $l=1$,定理 2.11 为一阶导数估计量提供了几乎确定一致收敛速度

$$\begin{aligned}|\nabla^{(1)}\hat{g}(x)-\nabla^{(1)}g(x)|&\equiv|\hat{g}^{(1)}(x)-g^{(1)}(x)|\\&=O((\ln(n)/(nh^{q+2}))^{1/2}+h^{p+1})\end{aligned}$$

类似地,人们从定理 2.11 中能够得到高阶导数估计量的几乎确定一致速度。

在这章中我们假定数据 $\{X_i,Y_i\}_{i=1}^{n}$ 是观测无误差的。然而,在实践中数据可能被污染或测量有误差。在非参数的背景中如何处理测量误差有丰富的文献。不过涵盖大量的这类文献超出了这本书的范围,我们推荐感兴趣的读者参考 Fan 和 Truong(1993),Carroll 和 Hall(2004),以及 Carroll,Maca 和 Ruppert(1999)及其参考文献。

92

2.6 应 用

2.6.1 职业声望数据

我们考虑以下使用来自 Fox(2002)在 R 的汽车图书馆数据的例子(R 开发核心队伍(2006))。数据集由 102 个观测值构成,每一个对应于一个特殊的职业。因变量是加拿大的职业声望,来自在 20 世纪 60 年代中期进行的一项社会调查的 Pineo-Porter 职业声望得分。解释变量是用 1971 年美元测量的每个职业

的平均收入。图 2.1 画出了这些数据和五个窗宽不同的局部线性回归估计，窗宽分别为低度平滑，过度平滑，Ruppert，Sheater 和 Wand(1995)的直接插入，Hurvich 等(1998)的修正 AIC(AIC_c)，以及交错鉴定法(Li 和 Racine(2004a))。始终使用一个二阶高斯核。

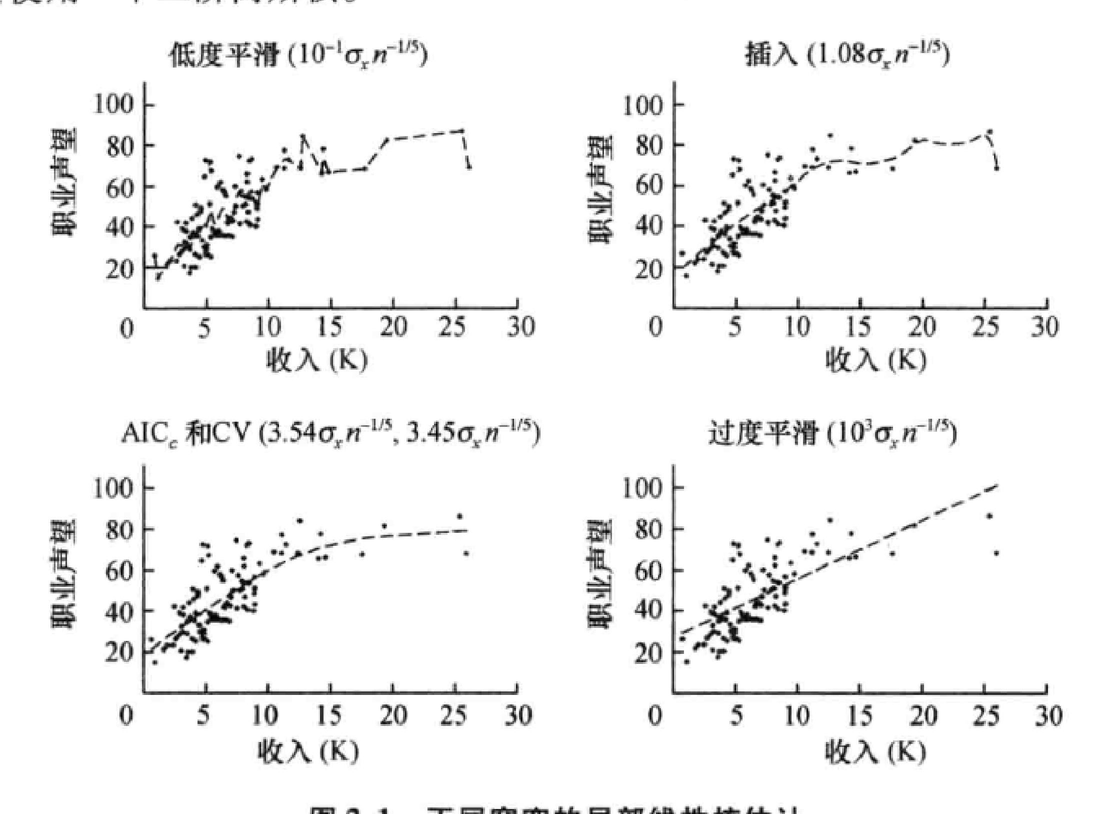


图 2.1 不同窗宽的局部线性核估计

注：AIC_c和 CV 窗宽是几乎相同的，因此在左下方的图中两条线看起来像一条。

在图 2.1 中过度平滑的局部线性估计是全局线性的，并且实际上就是所期望的 Y 关于 X 的简单线性最小二乘回归，而 AIC_c和 CV 准则看起来为这些数据提供了最合理的拟合。

2.6.2 青少年成长

在 1.13.3 节，我们提到青少年不正常的成长能够给孩子有某种医学上的问题提供一个早期的警告。我们使用取自疾病控制和预防中心(CDC)国家健康和营养调查关于健康美国孩子的人口数据来建立不同性别的身高和体重的联合分布模型。

我们现在使用即将在第 4 章简略介绍的混合数据局部线性估计量为“年龄身高均值”建模，考虑对男性和女性身高和年龄的潜在关系的回归函数。这个局部线性估计量使用的是根据 Hurvich 等(1998)的修正 AIC(AIC_c)准则选择的窗宽并使用一个二阶高斯核。年龄的窗宽是 7.63，而性别的窗宽是 0。 93

图 2.2 给出了按年龄和性别的平均身高。有趣的是,注意直到大约 10—12 岁,不同性别的平均身高在视觉上是很难区分的,然后则有明显的差异。在参数背景下,如果没有对样本进行划分,这种情况将特别难以建模,而非参数方法可以很好地揭示这种差别。

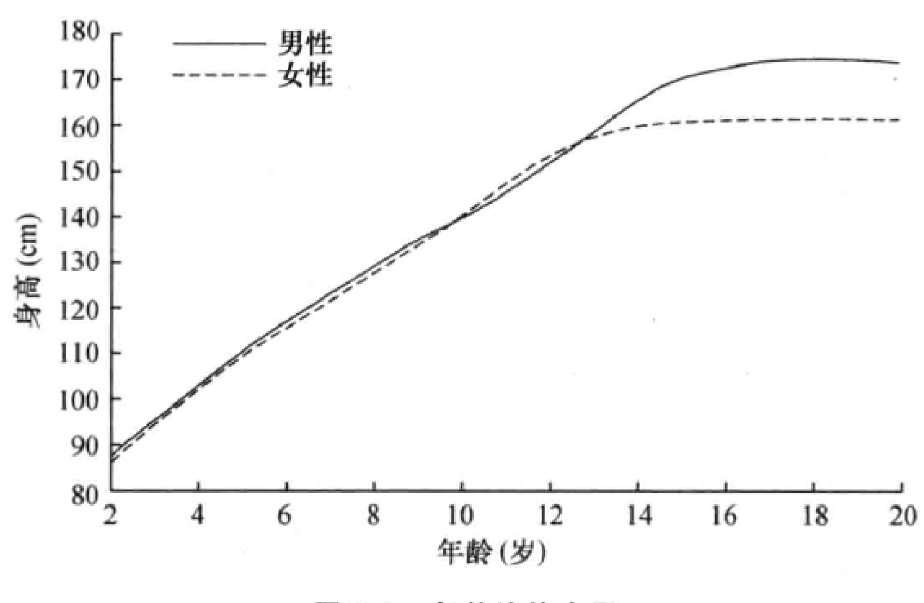


图 2.2 年龄均值水平

2.6.3 通货膨胀预测和货币增长

94

在货币领域中盛行的传统知识是货币增长应该对通货膨胀有预测力。然而,实证结果压倒性地指出恰恰相反,也就是说,货币增长对预测通货膨胀无论如何都没有预测力。而且,这个发现对样本期的变化和不同的计量方法都是稳健的(见 Leeper 和 Roush(2003),Stock 和 Watson(1999))。这个理论和经验的不一致在宏观经济学中是一个很著名的谜题。这个领域的许多实证研究是集中于使用如下形式的线性向量自回归(VAR)模型得到的预测

$$X_t = \alpha + \beta(L)X_{t-1} + \varepsilon_t$$

其中 $X_t = (\pi_t, Z_t)'$,π_t 是 t 时的通货膨胀率,且 Z_t 是货币总量的增长率。

Bachmeier, Leelahanon 和 Li(2007)使用从 1959 年 1 月到 2002 年 5 月的月度数据,基于非参数的视角重新研究了这个问题。使用的货币总量是 M1、M2 和 M3,以及相应的 M1、M2 和 M3 的 Divisia 金融服务指数,其中通货膨胀使用消费者价格指数来衡量。

使用来自 Bachmeier 等(2007)的数据,我们首先考虑使用一个关于通货膨胀和货币增长的二元 VAR 模型的参数通货膨胀预测,随着数据量的增加,它使用了一种递归估计方法;每做出一个预测是基于这样一个模型,它的估计使用

了所有可用的数据，包括已经做出预测的数据。对每一个模型以及预测区间(1994年1月到2002年4月)都进行预测。对 $s=1,6$ 和12个月的预测区间每一个模型的均方预测误差(MSPE)都被计算。使用SIC最优滞后长度选择得到通货膨胀和经济增长的两期滞后值，从而预测方程是 95

$$\pi_t = \alpha_0 + \alpha_1\pi_{t-s} + \alpha_2\pi_{t-s-1} + \alpha_3\Delta m_{t-s} + \alpha_4\Delta m_{t-s-1} + \varepsilon_t$$

其中 m_t 可以是这六种货币总量指标($\Delta m_t = m_t - m_{t-1}$)的任何一个。因为稳健性的缘故，我们报告了货币总量的一期和二期滞后值的结果，并且表2.1和表2.2报告了VAR模型两种情况(包括/排除每个滞后一期和二期的货币总量)下的MSPE比率；小于1.00的值意味着对一个给定的预测期包括此滞后货币总量改进了预测。

表2.1 相对参数均方预测误差(包括/排除每个滞后一期的货币总量)

区间	M1	M2	M3	M1D	M2D	M3D
1个月	1.05	1.02	1.06	1.00	1.00	1.00
6个月	0.90	1.00	1.11	1.00	1.04	1.06
12个月	0.93	0.96	1.09	1.00	1.02	1.04

表2.2 相对参数均方预测误差(包括/排除每个滞后二期的货币总量)

区间	M1	M2	M3	M1D	M2D	M3D
1个月	1.04	0.99	1.04	1.00	1.00	1.00
6个月	0.91	1.00	1.10	0.99	1.04	1.07
12个月	0.92	0.94	1.06	1.00	1.01	1.05

表2.1和表2.2显示通过包括货币作为一个预测元没有明显的系统改进。对包括货币总量的一期滞后，9/18预测更差，3/18改进，6/18没有变化。对二期滞后，8/18更差，5/18改进，5/18没有变化。总体来看，使用货币增长作为通货膨胀的预测在参数模型中可以被看作是相当弱的，因为包括货币总量导致更低效率的预测。 96

下一步我们把一个参数AR(2)模型的预测和一个如下形式的局部常数非参数AR(2)模型的预测进行比较

$$\pi_t = g(\pi_{t-s}, \pi_{t-s-1}) + \varepsilon_t$$

其中 $g(\cdot)$ 是未知的。我们对估计样本采用去一交错鉴定法，然后使用交错鉴定窗宽来生成我们的样本外预测。[①] 表2.3报告了非参数AR(2)模型的相

① 注意Bachmeier等(2007)使用了一个更复杂的两步交错鉴定程序。因此，这里报告的结果与在Bachmeier等(2007)中报告的结果有微小的不同，因为使用了不同的窗宽选择程序。

对均方预测误差以及参数 AR(2)模型的均方预测误差。

表 2.3　非参数 AR(2)/参数 AR(2) 的相对均方预测误差

区间	
1 个月	0.98
6 个月	1.01
12 个月	0.88

表 2.3 说明 1 个月和 6 个月期限的交错鉴定局部常数 AR(2)模型跟参数模型的预测能力不相上下，但与 12 个月期限的参数模型相比，非参数模型的预测有较大改进。由于非参数模型比一个正确设定的参数模型缺乏效率，这意味着 12 个月期限的参数模型的错误设定。然后人们会怀疑关于货币总量的预测无效率的结论实际上是参数错误设定的一个典型产物。我们因此考虑一个如下形式的局部常数非参数模型

$$\pi_t = g(\pi_{t-s}, \pi_{t-s-1}, \Delta m_{t-s}, \Delta m_{t-s-1}) + \varepsilon_t \tag{2.73}$$

其中 $g(\cdot)$是未知的，并且我们对窗宽选择仍使用去一交错鉴定法。

表 2.4 和表 2.5 给出了非参数模型的相对均方预测误差，有包括货币总量和不包括货币总量的。这些表格显示当包括滞后一期的货币总量时，2/18 的预测更差（在参数模型中是 9/18），9/18 改进（在参数模型中是 3/18），7/18 没有变化（在参数模型中是 6/18）；当包括滞后二期时，3/18 的预测更差，8/18 改进，7/18 没有变化。在一个非参数模型中，使用货币增长作为通货膨胀的一个预测变量在事实上是非常有帮助的，因为总体而言包括货币总量导致更准确的预测。

97

表 2.4　相对非参数的均方预测误差（包括/不包括滞后一期的货币总量）

区间	M1	M2	M3	M1D	M2D	M3D
1 个月	1.07	1.00	0.85	0.92	0.89	0.89
6 个月	0.96	0.95	1.15	1.00	1.00	1.00
12 个月	0.96	0.84	0.98	1.00	1.00	1.00

表 2.5　相对非参数的均方预测误差（包括/不包括滞后二期的货币总量）

区间	M1	M2	M3	M1D	M2D	M3D
1 个月	1.07	0.98	1.06	1.00	1.00	0.88
6 个月	1.00	0.96	1.07	1.00	1.00	1.00
12 个月	0.99	0.83	0.98	1.00	0.97	0.93

所以，稳健的非参数方法意味着货币增长非线性地影响通货膨胀。因此，前面的那个谜看起来是来自使用了误设线性关系的一个典型产物。

2.7 证 明

在这一部分我们提供本章前文中被省略的一些证明。

98

2.7.1 式(2.24)的推导

由于我们仅仅考虑所有回归元相关时的情况，我们因此假定

$$(h_1,\cdots,h_q) \in H_n = \{(h_1,\cdots,h_q) \mid (h_1,\cdots,h_q) \in [0,\eta_n]^q,\ nh_1\cdots h_q \geqslant t_n\} \tag{2.74}$$

其中 η_n 是一个正的序列，其收敛于零的速度比任何一个多项式对 n 的倒数更慢，并且 t_n 是一个发散到无穷的常数序列。方程式(2.74)基本要求 $nh_1\cdots h_q \to \infty$ 且 $h_s \to 0(s=1,\cdots,q)$。

令$\mathcal{S}$代表 ω 的支撑。我们还假定

$g(\cdot)$、$f(\cdot)$和 σ^2 有二阶连续导数；ω 是连续、非负并且有紧支撑的；$f(\cdot)$ 对 $x\in\mathcal{S}$远离零时是有界的。 (2.75)

使用式(2.23)的 $g_i=g(X_i)$，$\hat{g}_i=\hat{g}_{-i}(X_i)$，且 $M_i=M(X_i)$，我们有

$$\mathrm{CV}_{lc}(h_1,\cdots,h_q) = n^{-1}\sum_{i=1}^{n}(g_i-\hat{g}_i)^2 M_i + 2n^{-1}\sum_{i=1}^{n}u_i(g_i-\hat{g}_i)M_i + n^{-1}\sum_{i=1}^{n}u_i^2 M_i \tag{2.76}$$

式(2.76)右边的第三项不依赖于$(h_1\cdots h_q)$。可以证明第二项有一个比第一项更小的阶(见习题2.3)。因此，渐近地，最小化 $\mathrm{CV}_{lc}(h)$等价于最小化 $\sum_{i=1}^{n}[g_i-\hat{g}_i]^2 M_i$，式(2.76)右边的第一项。

我们接下来分析 CV_{lc}的首项。注意

$$\begin{aligned}\hat{g}_i - g_i &= (\hat{g}_i - g_i)\hat{f}_i/\hat{f}_i \\ &= (\hat{g}_i - g_i)\hat{f}_i/f_i + (\mathrm{s.o.}) \\ &\equiv \hat{m}_i/f_i + (\mathrm{s.o.})\end{aligned}$$

其中 $\hat{m}_i=(\hat{g}_i-g_i)\hat{f}_i$。定义

$$\hat{m}_{1i} = (n-1)^{-1}\sum_{l\neq i} g_l K_{h,il} - g_i\hat{f}_i$$

并且

99

$$\hat{m}_{2i} = (n-1)^{-1} \sum_{l \neq i}^{n} u_l K_{h,il}$$

其中 $K_{h,ij} = \prod_{s=1}^{q} h_s^{-1} k((x_{is} - x_{js})/h_s)$。那么 $\hat{m}_i = \hat{m}_{1i} + \hat{m}_{2i}$。因此，$\hat{m}_i \hat{f}_i^{-1} = (\hat{m}_{1i} + \hat{m}_{2i}) f_i^{-1}$ 是 $\hat{g}_i - g_i$ 的首项。因此，$\mathrm{CV}_{lc}(h)$ 的首项是 $\mathrm{CV}_{lc,0}(h) = n^{-1} \sum_{i=1}^{n} [\hat{m}_{1i} + \hat{m}_{2i}]^2 f_i^{-2} M_i$。可以进一步证明 $\mathrm{CV}_{lc,0}(h) = E[\mathrm{CV}_{lc,0}(h)] + (\mathrm{s.o.})$。因此，$E[\mathrm{CV}_{lc,0}(h)]$ 是 $\mathrm{CV}_{lc}(h)$ 的首项。现在，

$$\begin{aligned} E[\mathrm{CV}_{lc,0}(h_1, \cdots, h_q)] &= E\{[(\hat{m}_{1i} + \hat{m}_{2i})^2 f_i^{-2}] M_i\} \\ &= E[\hat{M}_{1i}^2 f_i^{-2} M_i] + E[\hat{m}_{2i}^2 f_i^{-2} M_i] \end{aligned} \tag{2.77}$$

其中我们已经使用 $E[\hat{m}_{1i}\hat{m}_{2i} f_i^{-2} M_i] = 0$（由于 $E(u_i | \{X_j\}_{j=1}^{n}) = 0$）。那么

$$\begin{aligned} E[\hat{m}_{1i}^2 f_i^{-2} \mid X_i] &= E[(g_j - g_i) K_{h,ij} (g_l - g_i) K_{h,il} f_i^{-2} \mid X_i] \\ &\quad + n^{-1} E[(g_j - g_i)^2 K_{h,ij}^2 f_i^{-2} \mid X_i] \\ &= \{E[(g_j - g_i) K_{h,ij} \mid x_i] f_i^{-1}\}^2 \\ &\quad + O\left((nh_1 \cdots h_q)^{-1} \sum_{s=1}^{q} h_s^2\right) \\ &= \left\{\frac{1}{f(X_i)} \int (g(X_i + hv) - g(X_i) K(v) f(X_i + hv) \mathrm{d}v)\right\}^2 + O(\eta_1 \eta_2) \\ &= \left\{\sum_{s=1}^{q} B_s(X_i) h_s^2\right\}^2 + O(\eta_2^3 + \eta_1 \eta_2) \end{aligned} \tag{2.78}$$

其中 $B_s(x)$ 跟式(2.8)中的定义是相同的，$\eta_1 = (nh_1 \cdots h_{q_1})^{-1}$ 且 $\eta_2 = \sum_{s=1}^{q_1} h_s^2$。

使用式(2.78)我们有

$$\begin{aligned} E(\hat{m}_{1i}^2 f_i^{-2} M_i) &= \int \left\{\sum_{s=1}^{q} B_s(x) h_s^2\right\}^2 f(x) M(x) \mathrm{d}x \\ &\quad + O(\eta_2^3 + \eta_1 \eta_2) \end{aligned} \tag{2.79}$$

100 接下来，

$$\begin{aligned} E[\hat{m}_{2i}^2 f_i^{-2} M_i] &= E\{f_i^{-2} M_i E[\hat{m}_{2i}^2 \mid X_i]\} \\ &= n^{-1} E\{f_i^{-2} M_i E[u_j^2 K_{h,ij}^2 \mid X_i]\} \\ &= (nh_1 \cdots h_q)^{-1} \\ &\quad \times E\left\{f_i^{-2} M_i \int f(X_i + hv) \sigma^2(X_i + hv) K^2(v) \mathrm{d}v\right\} \\ &= (nh_1 \cdots h_q)^{-1} \end{aligned}$$

$$\times E\{f_i^{-2}M_i[\kappa^q f(X_i)\sigma^2(X_i)+O(\eta_2)]\}$$

$$=\frac{\kappa^q}{nh_1\cdots h_q}\int\sigma^2(x)M(x)\mathrm{d}x+O(\eta_1\eta_2) \tag{2.80}$$

把式(2.79)和式(2.80)代入式(2.77)我们得到

$$E[\mathrm{CV}_{lc,0}(h_1,\cdots,h_q)]=\int\left\{\sum_{s=1}^{q}B_s(x)h_s^2\right\}^2 f(x)M(x)\mathrm{d}x$$

$$+\frac{\kappa^q}{nh_1\cdots h_q}\int\sigma^2(x)M(x)\mathrm{d}x+o(\eta_2^2+\eta_1)$$

$$=n^{-4(q+4)}\chi(a_1,\cdots,a_q)+o(n^{-4/(q+4)}) \tag{2.81}$$

其中 $a_s s$ 是通过 $h_s=a_s n^{-1/(q+4)}$ $(s=1,\cdots,q)$ 被定义的，且

$$\chi(a_1,\cdots,a_q)=\int\left\{\sum_{s=1}^{q}B_s(x)a_s^2\right\}^2 f(x)M(x)\mathrm{d}x$$

$$+\frac{\kappa^q}{a_1\cdots a_q}\int\sigma^2(x)M(x)\mathrm{d}x \tag{2.82}$$

2.7.2 定理 2.7 的证明

我们简述定理 2.7 证明的一个概要，而把具体细节留作习题。

从处理式(2.50)中产生的一个问题是

$$\frac{1}{n}\sum_i K_h(X_i,x)\begin{bmatrix}1\\X_i-x\end{bmatrix}(1,(X_i-x)')$$

是一个(渐近的)奇异矩阵，因此是不可逆的。我们可以通过在式(2.50)中间插入一个单位矩阵 $I_{q+1}=G_n^{-1}G_n$ 以一种等价的形式重写式(2.50)，并且 $G_n=\begin{bmatrix}1, & 0\\0, & D_h^{-2}\end{bmatrix}$，其中 D_h^{-2} 是一个 $q\times q$ 对角矩阵，其第 s 个对角元素由 h_s^{-2} 给出，我们得到(应用 $B^{-1}C=B^{-1}G_n^{-1}G_nC=(G_nB)^{-1}G_nC$)

101

$$\hat{\delta}(x)=\left[\sum_i K_h(X_i,x)G_n\left[\begin{matrix}1\\X_i-x\end{matrix}(1,(X_i-x)')\right]\right]^{-1}$$

$$\times\sum_i K_h(X_i,x)G_n\begin{bmatrix}1\\X_i-x\end{bmatrix}y_i$$

$$=\left[\sum_i K_h(X_i,x)\begin{bmatrix}1\\D_h^{-2}(X_i-x)\end{bmatrix}(1,(X_i-x)')\right]^{-1}$$

$$\times\sum_i K_h(X_i,x)\begin{bmatrix}1\\D_h^{-2}(X_i-x)\end{bmatrix}y_i \tag{2.83}$$

在证明中使用式(2.83)的优点是

$$\frac{1}{nh_1\cdots h_q}\sum_i K_h(X_i,x)\begin{bmatrix}1\\ D_h^{-2}(X_i-x)\end{bmatrix}(1,(X_i-x)')$$

以概率收敛于非奇异矩阵。因此,我们能够分别分析式(2.83)的分子和分母并因此极大地简化分析。

使用由下式给出的一个 Taylor 展开

$$\begin{aligned}g(X_i)&=g(x)+(X_i-x)'g^{(1)}(x)+(X_i-x)'g^{(2)}(x)(X_i-x)/2+R_m(x,X_i)\\&=(1,(X_i-x)')\delta(x)+(X_i-x)'g^{(2)}(x)(X_i-x)/2+R_m(x,X_i)\end{aligned}$$

102 其中 $R_m(x,X_i)$ 是余项,我们有

$$\begin{aligned}\hat{\delta}(x)&=\left[\frac{1}{n}\sum_i K_{h,ix}\begin{bmatrix}1\\ D_h^{-2}(X_i-x)\end{bmatrix}(1,(X_i-x)')\right]^{-1}\\&\quad\times\left\{\frac{1}{n}\sum_i K_{h,ix}\begin{bmatrix}1\\ D_h^{-2}(X_i-x)\end{bmatrix}[g(X_i)+\varepsilon_i]\right\}\\&=\delta(x)+\left[\frac{1}{n}\sum_i K_{h,ix}\begin{bmatrix}1, & (X_i-x)'\\ D_h^{-2}(X_i-x), & D_h^{-2}(X_i-x)(X_i-x)'\end{bmatrix}\right]^{-1}\\&\quad\times\left\{\frac{1}{n}\sum_i K_{h,ix}\begin{bmatrix}1\\ D_h^{-2}(X_i-x)\end{bmatrix}\right.\\&\quad\left.\times[(X_i-x)'g^{(2)}(x)(X_i-x)/2+\varepsilon_i+R_m(x,X_i)]\right\}\\&\equiv\delta(x)+[A^{1,x}]^{-1}\{A^{2,x}+A^{3,x}\}+(\text{s. o.})\end{aligned}\tag{2.84}$$

其中

$$A^{1,x}=\frac{1}{n}\sum_i K_{h,ix}\begin{bmatrix}1, & (X_i-x)'\\ D_h^{-2}(X_i-x), & D_h^{-2}(X_i-x)(X_i-x)'\end{bmatrix}\tag{2.85}$$

$$A^{2,x}=\frac{1}{2n}\sum_i K_{h,ix}\begin{bmatrix}1\\ D_h^{-2}(X_i-x)\end{bmatrix}(X_i-x)'g^{(2)}(x)(X_i-x)\tag{2.86}$$

$$A^{3,x}=\frac{1}{n}\sum_i K_{h,ix}\begin{bmatrix}1\\ D_h^{-2}(X_i-x)\end{bmatrix}\varepsilon_i\tag{2.87}$$

且低阶(s. o.)项来自

$$[A^{1,x}]^{-1}\frac{1}{n}\sum_i K_{h,ix}R_m(x,X_i)$$

它有一个比 $[A^{1,x}]^{-1}A^{2,x}$ 更低的阶。

习题 2.7 证明了 $A^{1,x}=Q+o_p(1)$，其中

$$Q=\begin{bmatrix} f(x), & 0 \\ \kappa_2 f^{(1)}(x), & \kappa_2 f(x) I_q \end{bmatrix}$$

使用分块矩阵的逆，我们得到

$$Q^{-1}=\begin{bmatrix} 1/f(x), & 0 \\ -f^{(1)}(x)/f^2(x), & I_q/(\kappa_2 f(x)) \end{bmatrix}$$

下一步，重写式(2.84)为

$$D(n)(\hat{\delta}(x)-\delta(x))=D(n)[A^{1,x}]^{-1}[A^{2,x}+A^{3,x}]+(\text{s. o.}) \quad (2.88)$$

并定义对角矩阵

$$R=\operatorname{diag}(Q^{-1})=\begin{bmatrix} 1/f(x), & 0 \\ 0, & I_q/(\kappa_2 f(x)) \end{bmatrix}$$

和

$$V=\begin{bmatrix} \kappa^q \sigma^2(x) f(x), & 0 \\ 0, & \kappa^{q-1}\kappa_{22}\sigma^2(x) f(x) I_q \end{bmatrix}$$

如果我们能够证明以下条件就可以证明定理 2.7：

(i) $D(n)[A^{1,x}]^{-1}[A^{2,x}+A^{3,x}]=D(n)Q^{-1}[A^{2,x}+A^{3,x}]+o_p(1)$。

(ii) $D(n)Q^{-1}[A^{2,x}+A^{3,x}]=RD(n)[A^{2,x}+A^{3,x}]+o_p(1)$。

(iii) $D(n)A^{2,x}=\begin{bmatrix} (nh_1\cdots h_q)^{1/2}(\kappa_2/2)f(x)\sum_{s=1}^{q} g_{ss}(x)h_s^2 \\ 0 \end{bmatrix}+o_p(1)$。

(iv) $D(n)A^{3,x}\to N(0,V)$ 依分布收敛。

(i) 到(iv)的证明在下面由四个引理给出。陈述(i)到(iv)意味着

$$D(n)\left[\hat{\delta}(x)-\delta(x)-\begin{bmatrix} (\kappa_2/2)f(x)\sum_{s=1} g_{ss}(x)h_s^2 \\ 0 \end{bmatrix}\right]$$

$$=RD(n)[A^{2,x}+A^{3,x}]-\begin{bmatrix} (nh_1\cdots h_q)^{1/2}(\kappa_2/2)\sum_{s=1}^{q} g_{ss}(x)h_s^2 \\ 0 \end{bmatrix}+o_p(1)$$

$$=RD(n)A^{3,x}+o_p(1)\to RN(0,V)+o_p(1)\to N(0,\Sigma)$$

依分布收敛，其中 $\Sigma=\mathrm{RVR}$，跟在定理 2.7 中给出的是相同的 Σ。

引理 2.1 $D(n)[A^{1,x}]^{-1}[A^{2,x}+A^{3,x}]=D(n)Q^{-1}[A^{2,x}+A^{3,x}]+o_p(1)$。

证明：记 $D(n)[A^{1,x}]^{-1}[A^{2,x}+A^{3,x}]=D(n)Q^{-1}[A^{2,x}+A^{3,x}]+D(n)[(A^{1,x})^{-1}-Q^{-1}][A^{2,x}+A^{3,x}]$，这足以证明

$$D(n)[(A^{1,x})^{-1}-Q^{-1}][A^{2,x}+A^{3,x}]=o_p(1) \tag{2.89}$$

104 从 $A^{1,x}=Q+o_p(1)$（见习题 2.7）我们知道

$$(A^{1,x})^{-1}=Q^{-1}+o_p(1) \tag{2.90}$$

令 $J\overset{\text{def}}{=}(A_{11}^{1,x})^{-1}-Q^{-1}\equiv\begin{bmatrix}J_{11}, & J_{12}\\ J_{21}, & J_{22}I_q\end{bmatrix}$。方程式(2.90)意味着对 $i,j=1,2$ 有 $J_{ij}=o_p(1)$。为证明式(2.89)，我们需要 J_{12} 有一个非常快的速度。以下我们证明 $J_{12}=O_p(\eta_2+\eta_1^{1/2}\eta_2^{1/2})$。

定义 $G=(A_{22}^{1,x}-A_{21}^{1,x}(A_{11}^{1,x})^{-1}A_{12}^{1,x})^{-1}$。使用分块矩阵求逆以及习题 2.7 的结果，我们得到

$$(A^{1,x})^{-1}=\begin{bmatrix}(A_{11}^{1,x})^{-1}(I_q+A_{12}^{1,x}GA_{21}^{1,x}(A_{11}^{1,x})^{-1}), & -(A_{11}^{1,x})^{-1}A_{12}^{1,x}G\\ -GA_{21}^{1,x}(A_{11}^{1,x})^{-1}, & G\end{bmatrix}$$

$$=\begin{bmatrix}(1/f(x))I_q+O_p(\eta_2), & O_p(\eta_2)\\ -f^{(1)}(x)/f(x)^2+o_p(1), & I_q/[\kappa_2 f(x)+o_p(1)]\end{bmatrix} \tag{2.91}$$

因为根据习题 2.7(ii)，有 $A_{12}^{1,x}=O_p(\eta_2+\eta_1^{1/2}\eta_2^{1/2})$。方程式(2.91)引出

$$J=(A^{1,x})^{-1}-Q^{-1}=\begin{bmatrix}O_p(\eta_2), & O_p(\eta_2)\\ o_p(1), & o_p(1)I_q\end{bmatrix} \tag{2.92}$$

这证明 $J_{12}=O_p(\eta_2)$。使用式(2.92)我们立刻得到

$$D(n)[(A^{1,x})^{-1}-Q^{-1}][A^{2,x}+A^{3,x}]=(nh_1\cdots h_q)^{1/2}\begin{bmatrix}1, & 0\\ 0, & D_h\end{bmatrix}$$

$$\times\begin{bmatrix}O_p(\eta_2), & O_p(\eta_2)\\ o_p(1), & o_p(1)I_q\end{bmatrix}\left[\begin{bmatrix}A_1^{2,x}\\ A_2^{2,x}\end{bmatrix}+\begin{bmatrix}A_1^{3,x}\\ A_2^{3,x}\end{bmatrix}\right]$$

$$=o_p(1)$$

因为根据习题 2.8 和 2.9，

$$(nh_1\cdots h_q)^{1/2}O_p(\eta_2)A_2^{2,x}=O_p((nh_1\cdots h_q)^{1/2}\eta_2^2)=o_p(1)$$

和

$$(nh_1\cdots h_q)^{1/2}O_p(\eta_2)A_2^{3,x}=(nh_1\cdots h_q)^{1/2}\eta_2 O_p((nh_1\cdots h_q h_s)^{-1/2})$$

$$=O_p(\eta_2^{1/2})$$

这证明了式(2.89)。 □

105 **引理 2.2** $D(n)Q^{-1}[A^{2,x}+A^{3,x}]=RD(n)[A^{2,x}+A^{3,x}]+o_p(1)$。

证明：注意 R 是一个对角矩阵，其中 $R=\text{diag}(Q^{-1})$，我们有

$$D(n)Q^{-1}[A^{2,x}+A^{3,x}]=RD(n)[A^{2,x}+A^{3,x}]+o_p(1)$$

因为与 Q^{-1} 的非对角元素相关的项全部是 $o_p(1)$ 的：

$$(nh_1\cdots h_q)^{1/2}h_sA_1^{2,x} = \sqrt{nh_1\cdots h_q}O_p(h_s\eta_2) = o_p(1)$$

并且

$$(nh_1\cdots h_q)^{1/2}h_sA_1^{3,x} = O_p(h_s) = o_p(1)$$

□

引理 2.3 $D(n)A^{2,x} = \begin{bmatrix} (nh_1\cdots h_q)^{1/2}(\kappa_2/2)f(x)\sum_{s=1}^{q}g_{ss}(x)h_s^2 \\ 0 \end{bmatrix} + o_p(1)$。

证明：根据习题 2.8，我们有

$$(nh_1\cdots h_q)^{1/2}A_1^{2,x} = \sqrt{nh_1\cdots h_q}(\kappa_2/2)f(x)\sum_{s=1}^{q}h_s^2g_{ss}(x) + o_p(1)$$

并且

$$(nh_1\cdots h_q)^{1/2}D_hA_2^{2,x} = \sqrt{nh_1\cdots h_q}O_p(\eta_2^{3/2}) = o_p(1)$$

因此

$$D(n)A^{2,x} = \begin{bmatrix} (nh_1\cdots h_q)^{1/2}A_1^{2,x} \\ (nh_1\cdots h_q)^{1/2}D_hA_2^{2,x} \end{bmatrix}$$

$$= \begin{bmatrix} (nh_1\cdots h_q)^{1/2}(\kappa_2/2)f(x)\sum_{s=1}^{q}g_{ss}(x)h_s^2 \\ 0 \end{bmatrix} + o_p(1)$$

□

引理 2.4 $D(n)A^{3,x} \to N(0,V)$ 依分布收敛，其中

$$V = \begin{bmatrix} \kappa^q f(x)\sigma^2(x), & 0 \\ 0, & \kappa^{q-1}\kappa_{22}f(x)\sigma^2(x)I_q \end{bmatrix}$$

证明：根据习题 2.9，

106

$$\mathrm{var}((nh_1\cdots h_q)^{1/2}A_1^{3,x}) = \kappa^q f(x)\sigma^2(x) + o(1),$$

$$\mathrm{var}((nh_1\cdots h_q)^{1/2}\cdot D_hA_2^{3,x}) = \kappa^{q-1}\kappa_{22}f(x)\sigma^2(x)I_q + o(1)$$

并且

$$\mathrm{cov}((nh_1\cdots h_q)^{1/2}A_1^{3,x},(nh_1\cdots h_q)^{1/2}D_hA_2^{3,x}) = o(1)$$

因此，$\mathrm{var}(D_{(n)}A^{3,x}) = V + o(1)$。同时注意 $A^{3,x}$ 有均值零，所以根据 Liapunov 中心极限定理，有 $D(n)A^{3,x}\xrightarrow{d}N(0,V)$。 □

2.7.3 在定理 2.10 中使用的 $A_{l,p+1}$ 和 V_l 的定义

我们使用在式(2.67)至式(2.69)中引入的符号来定义

$$\mu_j = \int v^jK(v)\mathrm{d}v, \quad \gamma_j = \int v^jK^2(v)\mathrm{d}v \tag{2.93}$$

$$M=\begin{bmatrix} M_{0,0} & M_{0,1} & \cdots & M_{0,p} \\ M_{1,0} & M_{1,1} & \cdots & M_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ M_{p,0} & M_{p,1} & \cdots & M_{p,p} \end{bmatrix},\quad \Gamma=\begin{bmatrix} \Gamma_{0,0} & \Gamma_{0,1} & \cdots & \Gamma_{0,p} \\ \Gamma_{1,0} & \Gamma_{1,1} & \cdots & \Gamma_{1,p} \\ \vdots & \vdots & \ddots & \vdots \\ \Gamma_{p,0} & \Gamma_{p,1} & \cdots & \Gamma_{p,p} \end{bmatrix} \tag{2.94}$$

其中 $M_{i,j}$ 和 $\Gamma_{i,j}$ 是 $N_i \times N_j$ 维的矩阵。例如，$M_{1,2}=\mu_{(1)} \times \mu_{(2)}$，其中 $\mu_{(1)}$ 和 $\mu_{(2)}$ 是 $N_1 \times 1$ 和 $N_2 \times 1$ 维的并且是从字典式排序的 μ_j 中定义的

$$\mu_{(1)}=\begin{bmatrix} \int v_1 K(v)\,\mathrm{d}v \\ \int v_2 K(v)\,\mathrm{d}v \\ \vdots \\ \int v_p K(v)\,\mathrm{d}v \end{bmatrix},\quad \mu_{(2)}=\begin{bmatrix} \int v_1^2 K(v)\,\mathrm{d}v \\ \int v_1 v_2 K(v)\,\mathrm{d}v \\ \vdots \\ \int v_p^2 K(v)\,\mathrm{d}v \end{bmatrix} \tag{2.95}$$

类似地，用 $K^2(\cdot)$ 取代 $K(\cdot)$，可以从 $M_{i,j}$ 中得到 $\Gamma_{i,j}$。

对两个 $m \times 1$ 维向量 C_1 和 C_2 我们使用 $C_1 \otimes C_2$ 来表示，它的第 j 个元素为 $C_{1,j}C_{2,j}(j=1,\cdots,m)$ 的一个 $m \times 1$ 维向量，也就是说，$\otimes$ 表示元素跟元素的乘积。

在定理 2.10 中引入的 V_l 矩阵是通过如下的 $N \times N$ 维矩阵 $V\left(N=\sum_{j=1}^{p} N_j\right)$ 定义的：

$$V=M^{-1}\Gamma M^{-1}$$

107 那么 V_l 是一个 $N_l \times N_l$ 维的矩阵，由从 $\sum_{j=0}^{l-1} N_j+1$ 到 $\sum_{j=0}^{l} N_j$ 的行与列构成。例如，V_0 是 V 的第一个对角元素。V_1 由 V 的从 2 到 $N_1+1=q+1$ 的行与列构成。V_2 由 V 的从 $N_0+N_1+1=q+2$ 到 $N_0+N_1+N_2=1+q+q(q+1)/2$ 的行与列构成，等等。

可以容易地核对当 $p=1$ 时（局部线性情况），我们有

$$M=\begin{bmatrix} 1, & 0 \\ 0, & \kappa_2 I_q \end{bmatrix}$$

$$\Gamma=\begin{bmatrix} \kappa^q, & 0 \\ 0, & \kappa^{q-1}\kappa_{22} I_q \end{bmatrix} \tag{2.96}$$

$$V=M^{-1}\Gamma M^{-1}=\begin{bmatrix} \kappa^q, & 0 \\ 0, & (\kappa^{q-1}\kappa_{ss}/\kappa_2^2) I_q \end{bmatrix}$$

其中 $\kappa_2=\int v^2 k(v)\,\mathrm{d}v$，$\kappa=\int k^2(v)\,\mathrm{d}v$，并且 $\kappa_{22}=\int v^2 k^2(v)\,\mathrm{d}v$。因此，$V_0=\kappa^q$ 且

$V_1 = (\kappa^{q-1}\kappa_{ss}/\kappa_2^2)I_q$。方差$[\sigma^2(x)/f(x)]V_l$对$l=0$和$l=1$准确地符合了定理2.7的结果。

定义一个$N_l \times 1$维向量$l_{(r)}$，其元素按字母的顺序随不同的$1/(r!) = 1/(r_1! \times r_2! \times \cdots \times r_q!)$，$\bar{r}=l$形式而变化。例如$2_j$是一个由下式给出的$N_2 \times 1$维向量

$$2_{(r)} = \begin{bmatrix} 1/2 \\ \iota_{q-1} \\ 1/2 \\ \iota_{q-2} \\ \vdots \\ 1/2 \end{bmatrix} \tag{2.97}$$

其中l_m是维数为m的向量。

然后定义

$$\mathcal{A} \equiv \begin{bmatrix} A_0 \\ A_1 \\ \vdots \\ A_p \end{bmatrix} \stackrel{\text{def}}{=} \begin{bmatrix} M_{0,p+1} \\ M_{1,p+1} \\ 2_{(r)\otimes M_{1,p+1}} \\ \vdots \\ p(r) \otimes M_{p,p+1} \end{bmatrix} \tag{2.98}$$

存在N_{p+1}个类似$(1/j!)(D^j g)(x)$的不同导数，其总阶数为$\bar{j}=p+1$：我们使用字典式排序把它们放在一个$N_{p+1} \times 1$维向量中，并且我们用$m_{p+1}(x)$来表示这个向量。例如，当$p=1$时（局部线性情况），我们有 108

$$m_{(2)}(x) \equiv \begin{bmatrix} \frac{1}{2}\frac{\partial^2 g(x)}{\partial x_1^2} \\ \frac{\partial^2 g(x)}{\partial x_1 \partial x_2} \\ \vdots \\ \frac{1}{2}\frac{\partial^2 g(x)}{\partial x_q^2} \end{bmatrix} \tag{2.99}$$

那么在定理2.10中引入的$A_{l,p+1}$是通过如下的方程式定义的($l=0,\cdots,p$)：

$$\mathcal{B} \stackrel{\text{def}}{=} M^{-1}\mathcal{A}m_{(p+1)}(x) \equiv \begin{bmatrix} A_{0,p+1} \\ A_{1,p+1} \\ \vdots \\ A_{p,p+1} \end{bmatrix} \tag{2.100}$$

其中 M 和 A 分别被定义于式(2.94)和式(2.98)，并且 $m_{p+1}(x)$ 是紧随式(2.98)被定义的。

当 $p=1$ 时(局部线性情况)，使用式(2.96)、式(2.99)和式(2.98)，我们有

$$\mathcal{B} = M^{-1}\mathcal{A}m_{(2)}(x) = \begin{bmatrix}(1/2)\kappa_2\mathrm{tr}[g^{(2)}(x)]\\ 0\end{bmatrix} \equiv \begin{bmatrix}A_{0,2}\\ A_{1,2}\end{bmatrix} \qquad (2.101)$$

使用以上在定理2.10中的 $A_{0,2}$ 和 $A_{1,2}$ 引出跟在定理2.7中相同的结果，当然它们应该就是这样。

2.8 习 题

习题 2.1 证明式(2.10)。

提示：注意

$$E\left[\hat{m}_1(x) - f(x)\sum_{s=1}^{q}B_s(x)h_s^2\right]^2 = \left\{E\left[\hat{m}_1(x) - f(x)\sum_{s=1}^{q}B_s(x)h_s^2\right]\right\}^2 + \mathrm{var}(\hat{m}_1(x))$$

然后应用式(2.8)和式(2.9)。

习题 2.2

(i) 对 $q=2$ 考虑式(2.82)：

$$\chi(a_1,a_2) = \int\left\{[a_1^2B_1(x) + a_2^2B_2(x)]^2 + \frac{\sigma^2(x)}{a_1a_2f(x)}\right\}f(x)M(x)\mathrm{d}x$$

推导最小化 $\chi(a_1,a_2)$ 的 a_1^0 和 a_2^0 的显性表达式。

109 (ii) 对一般多变量情况，χ_ω 可以被写作(见式(2.30))

$$\chi_z(z_1,z_2,\cdots,z_q) = z'Az + \frac{C_0}{\sqrt{z_1z_2\cdots z_q}}$$

其中 A 是一个 $q\times q$ 的正半定矩阵，且 $C_0>0$ 是一个正的常数。

证明当 A 是一个正半定矩阵时，那么 $\inf\chi_\nu>0$。这将意味着 $a_1^0,\cdots,a_q^0$ 全是正且有限的。

对(ii)的提示：这遵循在习题1.10中使用的相同的分析思路。

习题 2.3 证明

$$n^{-1}\sum_{i=1}^{n}u_i(g(X_i) - \hat{g}_{-i}(X_i))M(X_i) = o_p(\eta_2 + \eta_1^{1/2})$$

提示：记 $n^{-1}\sum_{i=1}^{n}u_i(g_i - \hat{g}_{-i})M_i = n^{-1}\sum_{i=1}^{n}u_i(g_i - \hat{g}_{-i})\hat{f}_iM_i/f_i + (\mathrm{s.o.})$，然后证明这个首项的二阶矩是 $O(n^{-1}(\eta_2^2+\eta_1))$。

习题 2.4 假定人们在交错鉴定方法中不使用去一估计量。

(i) 证明在这种情况下，对 n 的一个固定值，当 $h\to 0$ 时 $\mathrm{CV}(h)$ 逼近它的下界(0)，即 $\lim_{h\to 0+}\mathrm{CV}(h)=0$。也就是 $h\to 0$ 最小化了 $\mathrm{CV}(h)$。

(ii) 如果 h 任意接近于 0，简单地说，$\hat{g}(X_i,h=0)$，那么什么是相应的估计量？

我们知道 $h\to 0$ 太快违背了当 $n\to\infty$ 时 $nh_1\cdots h_q\to\infty$ 的条件。当 $n\to\infty$ 时这样一个估计量的方差趋向于无穷。

习题 2.5 这个问题解释了定义于式(2.39)的 $R(\bar{x},h_{q_1+1},\cdots,h_q)$。记 $\hat{g}(x)-\bar{g}(\bar{x})=[\hat{g}(x)-\bar{g}(\bar{x})]\hat{f}(x)/\hat{f}(x)=\hat{m}(x)/\hat{f}(x)$，其中 $\hat{m}(x)=\hat{m}_1(x)+\hat{m}_2(x)$，$\hat{m}_1(x)=n^{-1}\sum_{i=1}^{n}(\bar{g}_i-\bar{g}(\bar{x}))K_{h,xi}$，且 $\hat{m}_2(x)=n^{-1}\sum_{i=1}^{n}u_iK_{h,xi}$。记 $K_{h,xi}=\bar{K}_{h,xi}\tilde{K}_{h,xi}$，$\bar{K}_{h,xi}=\prod_{s=1}^{q_1}k((X_{is}-X_s)/h_s)$，并且 $\tilde{K}_{h,xi}=\prod_{s=q_1+1}^{q}k((X_{is}-X_s)/h_s)$，其中横线和波浪线分别指相关和不相关的变量。

定义 $\eta_{2,q_1}=\sum_{s=1}^{q_1}h_s^2$ 且 $\eta_{1,q_1}=(nh_1,\cdots,h_{q_1})^{-1}$。

(i) 证明

110

$$E[\hat{f}(x)]=\bar{f}(\bar{x})E[\tilde{K}_{h,xi}]+O(\eta_{2,q_1})$$

且

$$\mathrm{var}[\hat{f}(x)]=O(\eta_{1,q_1})=o(1)$$

因此，

$$\hat{f}(x)=\bar{f}(\bar{x})E[\tilde{K}_{h,xi}]+o_p(1)\tag{2.102}$$

注意式(2.102)意味着如下等式：

$$\hat{g}(x)-\bar{g}(\bar{x})=\frac{\hat{m}(x)}{\hat{f}(x)}=\frac{\hat{m}(x)}{\bar{f}(\bar{x})E[\tilde{K}_{h,xi}]}+(\mathrm{s.o.})\equiv A_n(x)+(\mathrm{s.o.})$$

其中 $A_n(x)=\hat{m}(x)/[\bar{f}(\bar{x})E(\tilde{K}_{h,xi})]$是 $\hat{g}(x)-\bar{g}(\bar{x})$的首项。

(ii) 证明

$$E[\hat{m}_1(x)]=\sum_{s=1}^{q_1}h_s^2\bar{B}_s(\bar{x})\bar{f}(\bar{x})E[\tilde{K}_{h,xi}]+o(\eta_{2,q_1}E(\tilde{K}_{h,xi}))$$

并且

$$\mathrm{var}[\hat{m}_1(x)]=O(\eta_{2,q_1}\eta_{1,q_1}[E(\tilde{K}_{h,xi})]^2)=o(\eta_{1,q_1}[E(\tilde{K}_{h,xi})]^2)$$

因此，

$$E[\hat{m}_1(x)^2] = \left\{ \sum_{s=1}^{q_1} h_s^2 \bar{B}_s(\bar{x})\bar{f}(\bar{x})E[\tilde{K}_{h,xi}] \right\}^2$$

$$+ o((\eta_{2,q_1}^2 + \eta_{1,q_1})[E(\tilde{K}_{h,xi})]^2) \qquad (2.103)$$

(iii) 证明

$$E[\hat{m}_2(x)^2] = \eta_{1,q_1}\kappa^{q_1}\bar{\sigma}^2(\bar{x})\bar{f}(\bar{x})E[\tilde{K}_{h,xi}^2][1+o(1)] \qquad (2.104)$$

(iv) 证明式(2.102)至式(2.104)意味着

$$E[A_n(x)^2] = [(\kappa_2/2)\sum_{s=1}^{q_1}\bar{B}_s(\bar{x})h_s^2]^2$$

$$+ \frac{\kappa^{q_1}\bar{\sigma}^2(\bar{x})}{nh_1\cdots h_{q_1}\bar{f}(\bar{x})}R(\bar{x},h_{q_1+1},\cdots,h_q) + (\text{s. o.}) \qquad (2.105)$$

其中 $R(\cdot)$是由式(2.39)定义的。

111 (v) 使用式(2.105)和

$$\text{CV} = \int E[\hat{g}(x) - \bar{g}(\bar{x})]^2 f(x)M(x)\mathrm{d}x \sim E[A_n^2(x)]$$

我们有

$$\text{CV} \sim \int \left\{ \left[\sum_{s=1}^{q_1}\bar{B}_s(\bar{x})h_j^2 \right]^2 + \right.$$

$$\left. \left[\frac{\kappa^{q_1}\bar{\sigma}^2(\bar{x})}{nh_1\cdots h_{q_1}\bar{f}(\bar{x})}\right]R(\bar{x},h_{q_1+1},\cdots,h_q) \right\} f(x)M(x)\mathrm{d}x \qquad (2.106)$$

证明 $h_s\to\infty$ $(s=q_1+1,\cdots,q)$最小化了式(2.106)。这将导致 $R(\cdot)=1$。使用 $R=1$ 证明式(2.106)导致式(2.35)。注意根据式(2.32)，$f(x)=\bar{f}(\bar{x})\tilde{f}(\bar{x})$。

基于以上的结果对定理 2.4 给出一个直观的证明。

习题 2.6

(i) 证明如果人们对所有 $s=1,\cdots,q$ 使用 $h_s=\infty$，那么局部线性估计量 $\hat{g}(x)$等同于最小二乘估计量 $\hat{\alpha}_0+x'\hat{\alpha}_1$，其中 $\hat{\alpha}_0$ 和 $\hat{\alpha}_1$ 是 α_0 和 α_1 基于 $Y_i=\alpha_0+X_i'\alpha_1$ + 误差项的最小二乘估计量(注意对这个问题我们不必假定真实回归函数 $g(\cdot)$对 x 是线性的)。

(ii) 证明当 $g(x)=\alpha_0+x'\alpha_1$ 对 x 是线性时，局部线性估计量是无偏的(注意在这个问题中你不应该假定 $h_s=\infty$)。

习题 2.7 令 $A_{ts}^{1,x}$ $(t,s=1,2)$如同在 2.7 节中定义的。证明如下结果：

(i) $A_{11}^{1,x}=f(x)+o_p(1)$。

(ii) $A_{12}^{1,x} = O_p\left(\sum_{s=1}^{q} h_s^2 + \left(\sum_{s=1}^{q} h_s\right)(nh_1\cdots h_q)^{-1/2}\right) = o_p(1)$。

(iii) $A_{21}^{1,x} = \kappa_2 f^{(1)}(x) + o_p(1)$。

(iv) $A_{22}^{1,x} = \kappa_2 f(x) I_q + o_p(1)$。

注意(i)—(iv)意味着

112

$$A^{1,x} = Q + o_p(1), Q = \begin{bmatrix} f(x), & 0 \\ \kappa_2 f^{(1)}(x), & \kappa_2 f(x) I_q \end{bmatrix}$$

提示:证明 $E[A^{1,x}] = Q + o(1)$ 以及 $\mathrm{var}[A^{1,x}] = o(1)$。

习题 2.8 令 $A_t^{2,x}(t=1,2)$ 如同在 2.7 节中定义的。证明如下结果:

(i) $A_1^{2,x} = (\kappa_2/2) f(x) \sum_{s=1}^{q} g_{ss}(x) h_s^2 + o_p(\eta_2)$。

(ii) $A_2^{2,x} = O_p(\eta_2)$。

提示:

(i) 证明 $E[A_1^{2,x}] = (\kappa_2/2) f(x) \sum_{s=1}^{q} g_{ss}(x) h_s^2 + o(\eta_2)$ 以及 $\mathrm{var}[A_1^{2,x}] = o(\eta_2^2)$。

(ii) 证明对所有 $s=1,\cdots,q, E[|A_{2,s}^{2,x}|] = O(\eta_2)$,其中 $A_{2,s}^{2,x}$ 是 $A_2^{2,x}$ 的第 s 个元素。

习题 2.9 令 $A_t^{3,x}(t=1,2)$ 如同在 2.7 节中定义的。证明如下结果($\eta_1 = (nh_1\cdots h_q)^{-1}$):

(i) $\mathrm{var}((nh_1\cdots h_q)^{1/2} A_1^{3,x}) = \kappa^q \sigma^2(x) f(x) + o(1)$。

(ii) $\mathrm{var}((nh_1\cdots h_q)^{1/2} D_h A_2^{3,x}) = \kappa_{22}\kappa^{q-1} f(x)\sigma^2(x) I_q + o(1)$。

(iii) $\mathrm{cov}((nh_1\cdots h_q)^{1/2} A_1^{3,x}, (nh_1\cdots h_q)^{1/2} D_h A_2^{3,x}) = O(\eta_2^{1/2}) = o(1)$。

注意以上(i)—(iii)意味着

$$\mathrm{var}(D(n)A^{3,x}) = \begin{bmatrix} \kappa^q \sigma^2(x) f(x), & 0 \\ 0, & \kappa_{22}\kappa^{q-1} f(x)\sigma^2(x) I_q \end{bmatrix} \equiv V$$

习题 2.10 证明当 $g(x)$ 在 x 是线性时($g(x) = a + x'b$),$g(x)$ 的局部线性估计量是 $g(x)$ 的一个无偏估计量。

注意:无偏性是一个有限样本概念(不是渐近性)。因此,人们不应该使用变量变换和 Taylor 展开方法来证明这个结果,因为这些方法假定 $h \to 0$ 且 $n \to \infty$。

习题 2.11 Pagan 和 Ullah(1999, pp. 154—155)提出了一种收入组合的(对数收入和年龄)局部常数核估计,它基于 $n = 205$ 个受过普通教育(高中)的男性

113

的加拿大数据(1971年加拿大人口普查公共使用版本)。而参数的二次型的设定是全局凹的,非参数拟合反映了在中位年龄(大约38岁)附近的一个"凹陷"。

(i) 计算并画出局部常数、局部线性和参数二次型的估计,对核估计量使用最小二乘交错鉴定法选择窗宽。

(ii) 那个凹陷出现在相应的非参数估计中了吗?

(iii) 使用标准误的渐近公式(即 $\hat{g}(x) \pm 2\hat{\sigma}(\hat{g}(x))$)沿着它们的误差边界画出非参数估计。不进行一个正式的检验,那个凹陷看起来是显著的吗?

(iv) 哪一个非参数估计量(即局部常数或者局部线性)看起来对这些数据提供了最"合适的"拟合?

第3章 混合数据的频率估计

传统的核方法典型地假定潜在数据在本质上是连续的。但遇到一个包括连续和离散数据的混合数据集时,传统非参数方法要根据对应离散变量的取值把数据划分为“单元”(cells),然后根据手中的问题使用每一个单元的数据来估计一个概率函数或一个回归函数。从现在起我们将把这种方法称作一种“基于频率”的方法。然而,理论和实践上均有很强的原因支持对一个连续和离散混合的数据类型应用核平滑方法,其中连续和离散的数据都被以一种特殊方式平滑。这种平滑方法,不同于下面介绍的频率方法,将在第4章中讨论。 115

为给在第4章中出现的材料打好基础,我们首先对严格离散变量的情况讨论概率函数和回归函数的传统的(即基于频率)非参数估计,然后我们在3.3节讨论基于频率的离散和连续变量混合的情况。在这一章中我们使用频率(非平滑)方法来处理离散变量。从理论角度来说,频率方法很容易处理,因为得到收敛速度和渐近正态性是非常直接的。然而,人们很快认识到,当样本规模相对于离散单元的数量不大时,频率方法在实践中是有问题的,这自然引出在第4章中要介绍的方法,其中我们展示了通过一种特殊的方式平滑离散变量,人们怎么能够经常得到比基于频率方法产生的更好的估计结果,特别是当单元数很多时。 116

3.1 离散数据的概率函数估计

当一个随机变量 X 以概率 $p(x)$ 取离散值时，函数 $p(x)$ 一般被称作"概率函数"(回忆我们把连续 X 的相应部分称作一个"概率密度函数"，以 $f(x)$ 表示)。使用上标 d 来表示一个离散变量，我们考虑一个 r 维离散随机变量 X_i^d，并且令 X_{is}^d 表示它的第 $s(s=1,\cdots,r)$ 个元素。假定 X_i^d 有有界支撑，那么，不失一般性，我们假定 X_{is}^d 的支撑是 $\{0,1,2,\cdots,c_s-1\}$，所以 X_i^d 的支撑是 $\mathcal{S}^d=\prod_{s=1}^{r}\{0,1,2,\cdots,c_s-1\}$，其中 $c_s\geq 2(s=1,\cdots,r)$ 是一个正整数。我们使用 $\mathbf{1}(\cdot)$ 来表示通常的示性函数($x^d\in\mathcal{S}^d$)，它假定$(\cdot)$为真时值为 1，否则为 0：

$$\mathbf{1}(X_i^d=x^d)=\begin{cases}1, & X_i^d=x^d\\ 0, & \text{其他}\end{cases}\tag{3.1}$$

那么，对任意 $x^d\in\mathcal{S}^d$，我们通过下式估计概率函数 $p(x^d)$

$$\tilde{p}(x^d)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i^d=x^d)\tag{3.2}$$

方程式(3.2)是 $p(x^d)$ 的标准和著名的"频率估计量"。注意如果我们把 $\mathbf{1}(X_i^d=x^d)$ 看作是一个非平滑核函数的话，这个估计量确实跟在式(1.8)中给出的连续数据的核估计量具有非常强的相似性。很容易看出

$$\begin{aligned}E[\tilde{p}(x^d)]&=E[\mathbf{1}(X_1^d=x^d)]\\&=\sum_{x_1^d\in\mathcal{S}^d}\mathbf{1}(x_1^d=x^d)p(x_1^d)\\&=p(x^d)\end{aligned}$$

117 因此，$\tilde{p}(x^d)$ 是无偏的。同样

$$\begin{aligned}\operatorname{var}(\tilde{p}(x^d))&=n^{-1}\operatorname{var}(\mathbf{1}(X_1^d=x^d))\\&=n^{-1}[E(\mathbf{1}(X_1^d=x^d))-(E\mathbf{1}(X_1^d=x^d))^2]\\&=n^{-1}[p(x^d)-p(x^d)^2]\\&=n^{-1}p(x^d)(1-p(x^d))\end{aligned}$$

其中我们使用了 $\mathbf{1}^2(\cdot)=\mathbf{1}(\cdot)$ 这个事实。因此，

$$\mathrm{MSE}(\tilde{p}(x^d))=n^{-1}p(x^d)(1-p(x^d))=O(n^{-1})$$

这意味着

$$\tilde{p}(x^d)-p(x^d)=O_p(n^{-1/2})\tag{3.3}$$

进一步，根据 Lindeberg-Levy 的中心极限定理，我们有

$$\sqrt{n}(\tilde{p}(x^d)-p(x^d))\xrightarrow{d}N(0,p(x^d)(1-p(x^d)))\tag{3.4}$$

注意以上的计算和我们在第 1 章中的均方误差(MSE)计算是一样的,当时我们考虑抛硬币的概率估计量($\hat{p}=P(H)$)。式(3.3)和式(3.4)表明,对包括离散变量的非参数估计,收敛速度比包括连续变量的非参数估计的情况更快。这个结果的原因是非常简单的,因为它来自我们的假定 x^d 的支撑是有限的。因此,对任何 $x^d \in \mathcal{S}^d$,我们有 $p(x^d)>0$,并且在这 n 个独立随机抽取的$X_1^d,\cdots,X_n^d$中,平均而言有 $np(x^d)=O(n)$ 个观测值,其取值为 x^d。因此,我们的估计量 $\hat{p}(x^d)$ 以$O_p((np(x^d))^{-1/2})=O_p(n^{-1/2})$的参数速度期望收敛于 $p(x^d)$。有限支撑假定意味着我们仅仅有有限多的待估参数($p(x^d), x^d \in \mathcal{S}^d$)。这在本质上是一个参数模型(因为当样本规模增加时参数的数量不增加)。因此,我们自然得到$\sqrt{n}$的速度,也就是标准的参数收敛速度。在统计学中也存在大量的文献来处理当样本规模增加时离散单元的数目也增加的情况,即使当 $n\to\infty$ 时,在每一个单元中观测值的数目也是有限的,这就是所谓的稀疏渐近性框架,我们推荐感兴趣的读者参考 Simonoff(1996)。在本书中我们不再进一步研究这种情况。

3.2 有离散回归元的回归

118

考虑一个由下式给出的只有离散回归元的非参数回归模型

$$Y_i = g(X_i^d) + u_i \tag{3.5}$$

其中 $X_i^d \in \mathcal{S}^d$,u_i 是独立同分布,其均值为零且 $E(u_i^2|X_i^d=x^d)=\sigma^2(x^d)$。我们考虑误差过程是条件异方差的并且形式未知。对任意 $x^d\in\mathcal{S}^d$,我们用下式估计 $g(x^d)$

$$\tilde{g}(x^d) = \frac{1}{n}\sum_{i=1}^{n} Y_i \mathbf{1}(X_i^d = x^d)/\tilde{p}(x^d) \tag{3.6}$$

其中 $\tilde{p}(x^d)$在式(3.2)中被定义。

容易证明

$$\tilde{g}(x^d) - g(x^d) = O_p(n^{-1/2}) \tag{3.7}$$

并且

$$\sqrt{n}(\tilde{g}(x^d) - g(x^d)) \xrightarrow{d} N(0, \sigma^2(x^d)/p(x^d)) \tag{3.8}$$

通过观察到 $\sum_{i=1}^{n}[g(X_i^d)-g(x^d)]\mathbf{1}(X_i^d=x^d)=0$。式(3.7)和式(3.8)的证明是相当直观的,留给读者作为习题(见习题 3.1)。

3.3 混合数据的估计:频率方法

我们现在把注意力转向离散和连续变量混合的情况。我们使用 X_i^c 来表示

一个连续变量，因此记 $X_i=(X_i^c,X_i^d)\in\mathbb{R}^q\times\mathcal{S}^d$。我们定义 $f(x)=(x^c,x^d)$ 是 (X_i^c,X_i^d) 的联合概率密度函数。

3.3.1 混合数据的密度估计

假定我们有取自 $f(\cdot)$ 的独立观测值 $X_1,X_2,\cdots,X_n$。$f(x)$ 的一个非参数核估计量由下式给出

$$\tilde{f}(x)=\frac{1}{n}\sum_{i=1}^{n}W_h(X_i^c,x^c)\mathbf{1}(X_i^d=x^d)\tag{3.9}$$

119 其中 $W_h=(X_i^c,x^c)=\prod_{s=1}^{q}h_s^{-1}\omega((X_{is}^c-x_s^c)/h_s)$，$\omega(\cdot)$ 是一个满足式(1.10)的标准单变量二阶核函数。我们在下一个定理中描述 $\tilde{f}(x)$ 的收敛速度和渐近分布。

定理 3.1 对所有 $x^d\in\mathcal{S}^d$，假定 $f(x)=(x^c,x^d)$ 满足跟在定理 1.3 中给出的 $f(x^c)$ 相同的条件。同时假定，当 $n\to\infty$ 时，$h_s\to 0$ $(s=1,\cdots,q)$，$n(h_1\cdots h_q)\left(\sum_{s=1}^{q}h_s^6\right)\to 0$，并且 $nh_1\cdots h_q\to\infty$。那么

(i) $\tilde{f}(x)-f(x)=O_p\left(\sum_{s=1}^{q}h_s^2+(nh_1\cdots h_q)^{-1/2}\right)$。

(ii) $(nh_1\cdots h_q)^{1/2}\left(\tilde{f}(x)-f(x)-(\kappa_2/2)\sum_{s=1}^{q}f_{ss}(x)h_s^2\right)$ 依分布收敛于 $N(0,\kappa^q f(x))$，其中 $f_{ss}(x)=\partial^2 f(x^c,x^d)/\partial(x_s^c)^2$ 是 $f(\cdot)$ 对 x_s^c 的二阶导数，并且其中 $\kappa_2=\int v^2\omega(v)\mathrm{d}v$ 且 $\kappa=\int\omega^2(v)\mathrm{d}v$。

定理 3.1 的证明在 3.5 节给出。从定理 3.1 中我们看到在混合变量的情况下 $\tilde{f}(x)$ 的收敛速度与在包括只有 q 个纯粹连续变量子集的情况下是相同的。这是个必然的结果，因为对纯粹离散变量的估计有一个收敛速度 $O_p(n^{-1/2})$，它比 q 个纯粹连续变量情况下的收敛速度更快。因此，在混合变量的情况下 $\tilde{f}(x)$ 的收敛速度是由两个速度中更慢的那个决定的，也就是由来自 q 个连续变量子集出现的收敛速度决定的。

3.3.2 混合数据的回归

以上的频率方法也可被用于估计有离散和连续回归元混合的回归函数。对 $x=(x^c,x^d)\in\mathbb{R}^q\times\mathcal{S}^d$，我们通过下式估计 $g(x)=E(Y|X=x)$

$$\tilde{g}(x)=n^{-1}\sum_{i=1}^{n}Y_iW_h(X_i^c,x^c)\mathbf{1}(X_i^d=x^d)/\tilde{f}(x)\tag{3.10}$$

其中 $\tilde{f}(x)$ 在式(3.9)中被定义。接下来的定理给出了 $\tilde{g}(x)$ 的渐近分布。

定理3.2 对任何固定的 $x^d \in \mathcal{S}^d$，假定 $g(x^c,x^d)$ 满足$\mathcal{S}$在定理2.2中给出的 $g(x^c)$ 相同的条件，并且也假定，当 $n\to\infty$ 时，$h_s\to 0(s=1,\cdots,q)$，$nh_1\cdots h_q\to\infty$ 且 $n(h_1\cdots h_q)\left(\sum_{s=1}^{q}h_s^6\right)\to 0$，那么，我们有

120

$$\sqrt{nh_1\cdots h_q}\left(\tilde{g}(x)-g(x)-\sum_{s=1}^{q}B_s(x)h_s^2\right)\xrightarrow{d}N(0,\kappa^q\sigma^2(x)/f(x))$$

其中 $B_s(x)=(\kappa_2/2)\{2f_s(x)g_s(x)+f(x)g_{ss}(x)\}/f(x)$ 与在定理2.2中的定义是相同的，$x=(x^c,x^d)$，并且 $g_s(x)(f_s(x))$ 和 $g_{ss}(x)$ 是 $g(f)$ 对 x_s^c 的一阶和二阶导数，其中 $\kappa_2=\int v^2\omega(v)\mathrm{d}v$ 且 $\kappa=\int\omega^2(v)\mathrm{d}v$。

这个证明类似于式(2.14)的证明(仅仅有纯粹连续变量的情况)，它被留作习题(见习题3.2)。

从定理3.2中我们再一次看到收敛速度是仅仅由跟连续变量有关的收敛速度来决定的。

沿用在第2章中的分析，人们能再次通过交错鉴定(CV)方法、广义交错鉴定方法，或者插入法选择平滑参数(h_s's)。在混合变量情况下为了使非参数频率估计方法产生可靠的结果，在每一个离散单元，人们需要一个适当的数据数量。如果离散单元的数量相对于样本规模偏大，非参数频率方法可能如同预期的那样是非常不可靠的。

3.4 关于频率方法一些要注意的说明

在这章我们已经证明，理论上对一个非参数估计框架加入离散变量(有有限支撑)是没有问题的，因为非参数频率方法(仅有离散变量时)有 $O_p(n^{-1/2})$ 的收敛速度，这比跟 q 个纯粹连续变量子集相关的收敛速度更快。然而，很明显的是，非参数频率方法仅仅在人们有一个大的样本并且离散变量取有限多的值时，即在离散单元的数目比样本规模小很多的情形下是有用的。然而，对于许多经济学数据集，人们经常遇到离散单元的数目接近于甚至大于样本数目的情况。显然，划分样本的频率估计量不能用于这样的情况。

为进一步解释这一点，令 M 代表跟离散变量有关的单元数量。即使在简单应用中，M 也可能是几百甚至几千。例如，如果人们有包括一个人的性别(0/1)、职业的行业(比如，1—20)，以及宗教信仰(比如，1—5)的数据，那么这个简单的出现三个离散变量的情况给出了 $2\times 20\times 5=200$ 个单元。在每个单元中观测值(有效样本大小)的平均数是 $n_{\text{eff}}=n/M$。由于 M 是一个固定有限数

121

（无论多大），渐近地，我们有 $n/M=O(n)$。因此，有效样本大小是 n 阶的。即使 $M=1\,000\,000$，n/M 仍然是 $O(n)$ 阶的。然而，对一个单变量连续变量，用于估计的有效样本大小是 $nh=O(n^{4/5})$ 阶的（如果 $h=O(n^{-1/5})$）。渐近地，对任何一个正常数 c 和 M，$cn^{4/5}$ 是比 n/M 更低的阶。因此，对离散变量的渐近有效样本规模大于对连续变量的情况。因此，有离散变量的非参数估计比连续变量的情形有更快的收敛速度。

然而在有限样本应用中，情况可能相当不同。就在小样本应用中的有效性而言，对离散变量的情况，$\sqrt{n}$ 的速度可能提供一个易误解的影响。让我们考虑一个有 5 个离散变量，样本大小为 $n=500$ 的一些人为的数据，并进一步假定这 5 个离散变量分别取 2，2，5，5 和 5 个不同值。在这个例子中，这 5 个离散变量出现产生的离散单元的数目是 $2\times2\times5\times5\times5=500$。因此，在每个单元中观测值的平均数（有效样本大小）是 $n/500=500/500=1$，这太小以至于不能产生任何有意义的非参数估计（即如果人们使用频率方法）。然而，为保证有一个连续变量的准确的非参数估计，$n=500$ 当然是足够大的。如果 $q=1$（$r=0$，即没有离散变量），那么对于有一个连续变量，在非参数估计中使用的样本大小是 $n^{4/5}=500^{4/5}\sim144$，这与以上在离散变量的例子中（单元的数目等于 500）有效样本大小为 1 相比，它要大得多。

在第 4 章中，我们讨论一种不同的方法，那里我们也使用了由 Aitchison 和 Aitken（1976）首先提出的一种特别方法来平滑离散变量。我们扩展了 Aitchison 和 Aitken 的方法来研究离散和连续变量混合的非参数回归和条件密度估计。我们还给出一些实证应用来说明，当跟一些经常使用的参数方法相比较时，这种平滑方法也能提供较好的样本外的预测。

122

3.5 证　明

3.5.1 定理 3.1 的证明

证明　对以下内容，我们使用下面的缩略符号

$$f(x^c+hv,x^d)=f(x_1^c+h_1v_1,\cdots,x_q^c+h_qv_q,x^d)$$

使用跟定理 1.3 的证明相同的推理，我们有

$$\begin{aligned}E[\tilde{f}(x)]&=E[W_h(X_1^c,x^c)\mathbf{1}(X_1^d=x^d)]\\&=(h_1,\cdots,h_q)^{-1}\sum_{x_1^d\in S^d}\int f(x_1^c,x_1^d)W\left(\frac{x_1^c-x^c}{h}\right)\mathbf{1}(x_1^d=x^d)\mathrm{d}x_1^c\\&=(h_1,\cdots,h_q)^{-1}\int f(x_1^c,x^d)W\left(\frac{x_1^c-x^c}{h}\right)\mathrm{d}x_1^c\end{aligned}$$

$$= \int f(x^c + hv, x^d) W(v) \mathrm{d}v$$

$$= f(x^c, x^d) + (\kappa_2/2) \sum_{s=1}^{q} f_{ss}(x) h_s^2 + O\left(\sum_{s=1}^{q} h_s^3\right)$$

下一步,通过跟定理 1.4 的证明相同的论证,我们有

123

$$\operatorname{var}(\tilde{f}(x)) = n^{-1}\operatorname{var}[W_h(X_1, x^c)\mathbf{1}(X_1^d = x^d)]$$

$$= (nh_1^2, \cdots, h_q^2)^{-1}\left\{E\left[W\left(\frac{X_1^c - x^c}{h}\right)^2 \mathbf{1}(X_1^d = x^d)\right]\right.$$

$$\left. - \left\{E\left[W\left(\frac{X_1^c - x^c}{h}\right)\mathbf{1}(X_1^d = x^d)\right]\right\}^2\right\}$$

$$= (nh_1^2, \cdots, h_q^2)^{-1}\left\{\sum_{x_1^d \in S^d}\int f(x_1^c, x_1^d)\right.$$

$$\times W\left(\frac{x_1^c - x^c}{h}\right)\mathbf{1}(x_1^d = x^d)\mathrm{d}x_1^c$$

$$\left. - \left[\sum_{x_1^d \in S^d}\int f(x_1^c, x_1^d) W\left(\frac{x_1^c - x^c}{h}\right)\mathbf{1}(x_1^d = x^d)\mathrm{d}x_1^c\right]^2\right\}$$

$$= (nh_1, \cdots, h_q)^{-1}\left\{f(x_1^c, x_1^d)\int W^2(v)\mathrm{d}v + O(h_1, \cdots, h_q)\right\}$$

$$\equiv (nh_1, \cdots, h_q)^{-1}\{\kappa^q f(x) + O(h_1, \cdots, h_q)\}$$

$$\left(由于\int W^2(v)\mathrm{d}v = \left[\int \omega^2(v)\mathrm{d}v\right]^q = \kappa^q\right)$$

以上结果证明了 $\operatorname{bias}(\tilde{f}(x)) = O\left(\sum_{s=1}^{q} h_s^2\right)$ 和 $\operatorname{var}(\tilde{f}(x)) = O((nh_1, \cdots, h_q)^{-1})$。因此,$\operatorname{MSE}(\tilde{f}(x)) = O\left(\left(\sum_{s=1}^{q} h_s^2\right)^2 + (nh_1, \cdots, h_q)^{-1}\right)$,这意味着定理 3.1(i) 成立。

(ii) 部分来自跟以上给出的同样的论证,应用 Liapunov 中心极限定理(见定理 1.3 的证明)即可证明定理 3.1(ii)。 □

3.6 习 题

习题 3.1 证明式(3.7)和式(3.8)。

提示:记

$$\bar{g}(x^d) - g(x^d) = (\bar{g}(x^d) - g(x^d))\bar{p}(x^d)/\bar{p}(x^d) \equiv \tilde{m}(x^d)/\bar{p}(x^d)$$

现在($Y_i = g(X_i^d) + u_i$),

124

$$\tilde{m}(x^d) = n^{-1}\sum_i [g(X_i^d) + u_i - g(x^d)]\mathbf{1}(X_i^d = x^d)$$
$$= n^{-1}\sum_i u_i\mathbf{1}(X_i^d = x^d) \qquad (3.11)$$

因为 $\sum_i [g(X_i^d) - g(x^d)]\mathbf{1}(X_i^d = x^d) = 0$。

从式(3.11)中,很容易看到 $E[\tilde{m}(x^d)] = 0$ 以及 $E[\tilde{m}(x^d)^2] = O(n^{-1})$,这意味着 $\tilde{m}(x^d) = O_p(n^{-1/2})$。

此外,$\tilde{p}(x^d) = p(x^d) + O_p(n^{-1/2})$。同时我们有 $\tilde{g}(x^d) - g(x^d) = O_p(n^{-1/2})$。

使用式(3.11),很容易看到 $E\{[\tilde{m}(x^d)]^2\} = n^{-1}\sigma^2(x^d)p(x^d)$。因此,根据 Lindeberg-Levy 中心极限定理,有 $n^{1/2}\tilde{m}(x^d) \xrightarrow{d} N(0, \sigma^2(x^d)p(x^d))$。因此,

$$\sqrt{n}(\tilde{g}(x^d) - g(x^d)) = \sqrt{n}\tilde{m}(x^d)/\tilde{p}(x^d) \xrightarrow{d} (1/p(x^d))N(0, \sigma^2(x^d)p(x^d))$$
$$= N(0, \sigma^2(x^d)/p(x^d))$$

习题 3.2 证明定理 3.2。

提示:类似式(2.14)的证明,记 $\tilde{g}(x) - g(x) = (\tilde{g}(x) - g(x))\tilde{f}(x)/\tilde{f}(x) \equiv \tilde{m}(x)/\tilde{f}(x)$。记 $\tilde{m}(x) = \tilde{m}_1(x) + \tilde{m}_2(x)$,其中 $\tilde{m}_1(x)$ 和 $\tilde{m}_2(x)$ 是与第 1 章的 $\tilde{m}_1(x)$ 和 $\tilde{m}_2(x)$ 用相同的方式定义的。证明

$$\tilde{m}_1(x) = (\kappa_2/2)\sum_{s=1}^{q}[f(x)g_{ss}(x) + 2f_s(x)g_s(x)]h_s^2$$
$$+ o_p\left(\sum_{s=1}^{q}h_s^2 + (nh_1\cdots h_q)^{-1/2}\right)$$

以及

$$(nh_1\cdots h_q)^{1/2}\tilde{m}_2(x) \to N(0, \kappa^q\sigma^2(x)f(x)) \text{ 依分布收敛}$$

这些结果和 $\tilde{f}(x) = f(x) + o_p(1)$ 推出定理 3.2。

第4章 混合数据的核估计

在第3章中，我们讨论了出现离散变量时，传统的基于频率方法的非参数估计，同时也概述了这类方法的优点和缺点。现在把我们的关注点转向一种不同的非参数方法，它可以运用于离散变量情形，这里我们以一种特别的方式来平滑离散变量，而不是在第3章中描述的使用传统的频率方法。 125

当估计概率函数和回归函数时，通过平滑离散变量而不是采用传统的基于单元的频率方法，我们可以大大地扩展非参数方法的范围。从统计学的角度看，平滑离散变量可能会引入偏误。然而，它也减小了有限样本方差，跟基于频率的估计量相比，非参数估计量的有限样本均方误差（MSE）会减小。

在4.5节中介绍了这种方法一个特别值得注意的特征，其中使用了平滑方法，并结合了数据驱动的窗宽选择方法，如交错鉴定法来自动去除"无关变量"（无关变量的详细定义在4.5节中给出）。这是一个重要的结论，因为在实际中我们已经注意到无关变量经常意外地出现。当无关变量存在时，在第3章介绍的传统的频率估计方法仍然把样本划分为许多离散单元，包括那些由于无关变量出现而产生的单元。相反，交错鉴定—平滑方法可以（渐近地）自动移除无关变量，从而把非参数模型的维数降低至只和相关变量有关。在这样的情形中，通过平滑离散变量常常得到得到令人印象深刻的效率。4.1节至4.3节讨论了只有离散变量的情况，4.4节和4.5节讨论了一般的离散和连续变量混合的情况。 126

4.1 离散数据联合分布的平滑估计

现在我们对定义在 $X^d \in \mathcal{S}^d$，X^d 的支撑上的离散数据，考虑一个概率函数的核估计。和以前一样，我们使用 x_s^d 和 X_{is}^d 来分别表示 x^d 和 $X_i^d(i=1,\cdots,n)$ 的第 s 个元素。参考 Aitchison 和 Aitken(1976)，对 $x_s^d, X_{is}^d \in \{0,1,\cdots,c_s-1\}$，我们定义一个离散一元核函数为：

$$l(X_{is}^d, x_s^d, \lambda_s) = \begin{cases} 1-\lambda_s, & X_{is}^d = x_s^d \\ \lambda_s/(c_s-1), & X_{is}^d \neq x_s^d \end{cases} \tag{4.1}$$

注意当 $\lambda_s=0$ 时，$l(X_{is}^d, x_s^d, 0) = \mathbf{1}(X_{is}^d = x_s^d)$ 变为示性函数，若 $\lambda_s=(c_s-1)/c_s$，则 $l\left(X_{is}^d, x_s^d, \frac{c_s-1}{c_s}\right) = 1/c_s$ 对于所有的 x_s^d 和 X_{is}^d 是一个常数。所以，λ_s 的值域为 $[0,(c_s-1)/c_s]$。

对于多元数据，我们使用一个下面给出的标准乘积核函数：

$$\begin{aligned} L(X_i^d, x^d, \lambda) &= \prod_{s=1}^{r} l(X_{is}^d, x_s^d, \lambda_s) \\ &= \prod_{s=1}^{r} \{\lambda_s/(c_s-1)\}^{N_{is}(x)}(1-\lambda_s)^{1-N_{is}(x)} \end{aligned} \tag{4.2}$$

其中 $N_{is}(x) = \mathbf{1}(X_{is}^d \neq x_s^d)$ 是一个示性函数，当 $X_{is} \neq x_s^d$ 时它等于 1，否则为 0。

对向量 $\lambda=(\lambda_1,\cdots,\lambda_r)$ 的一个给定值，我们通过下式来估计 $p(x^d)$：

$$\hat{p}(x^d) = \frac{1}{n}\sum_{i=1}^{n} L(X_i^d, x^d, \lambda) \tag{4.3}$$

127 注意式(4.1)中定义的权重函数 $l(.,.,.)$ 的和为 1，这保证了在式(4.3)中定义的 $\hat{p}(x^d)$ 是一个合适的概率指标，即 $\sum_{x^d \in \mathcal{S}^d} \hat{p}(x^d) = 1$。同样注意如果对所有的 $1 \leqslant s \leqslant r$，$\lambda_s=0$，那么式(4.3)将变为在第 3 章式(3.2)中定义的频率估计。因此，作为一种特别的情形，我们的平滑估计量 $\hat{p}(x^d)$ 嵌套了频率估计量 $\bar{p}(x^d)$。λ_s 取它值域的上界值 $(c_s-1)/c_s$ 时产生另外一种有趣的情况，此时，$\hat{p}(x^d)$ 与 x_s^d 不相关。在这种情况下我们说 x_s^d"被平滑出去了"(smoothed out)，于是，相应的估计量 $\hat{p}(x^d)$ 关于 x_s^d 是均匀分布，即对所有的 $x_s^d, z_s^d \in \mathcal{S}_s^d = \{0,1,\cdots,c_s-1\}$，有 $\hat{p}(x_{-s}^d, x_s^d) = \hat{p}(x_{-s}^d, z_s^d)$，其中 $x_{-s}^d = (x_1^d,\cdots,x_{s-1}^d,x_{s+1}^d,\cdots,x_r^d)$。如果对所有的 $x_{-s}^d \in \mathcal{S}^d/\mathcal{S}_s^d$（这里 $x \in A/B$ 是指 $x \in A$ 但 $x \notin B$），我们说 X^d 关于 x_s 是均匀分布的

$$p(x_{-s}^d, x_s^d) = p(x_{-s}^d, z_s^d)\text{，所有 } x_s^d, z_s^d \in \{0,1,\cdots,c_s-1\} \tag{4.4}$$

当估计一个联合概率密度函数时,使用平滑分类变量至少有两个理由。首先,尽管 $\lambda_s \neq 0$ 会导致一些估计偏误,但它减小了方差。因此,有限样本均方误差能减小。其次,当 X^d 关于 x_s 均匀分布时,$\lambda_s = (c_s - 1)/c_s$ 有效率地利用了这个信息,即它没有增加偏误但减小了方差。

以上的观点强调了数据驱动窗宽选择方法的可取之处,那就是如果 X^d 关于 x_s 是均匀分布的,则选择较大的 λ_s,反之则选择较小的 λ_s。这样的结果就是当 x_s^d 是均匀分布时,最小二乘交错鉴定法有很高的概率选择一个大的 λ_s 值。

在许多经济数据集中,离散单元的数量接近甚至超过样本的大小,使得频率估计方法不可靠或完全不可能应用。然而,在这种情况下,人们可以聪明地使用在式(4.3)中定义的平滑估计量,因为在式(4.3)中我们避免了当估计 $p(x^d)$ 时把数据划分为不同的单元。

假定当 $n \to \infty$ $(s = 1, \cdots, r)$ 时,$\lambda_s \to 0$,容易证明(见习题 4.3):　128

$$
\begin{aligned}
E[\hat{p}(x^d)] &= p(x^d) + \sum_{s=1}^{r} B_{p,s}\lambda_s + O\left(\sum_{s=1}^{r} \lambda_s^2\right) \\
\operatorname{var}(\hat{p}(x^d)) &= \frac{p(x^d)(1 - p(x^d))}{n} + O\left(n^{-1}\sum_{s=1}^{r} \lambda_s\right)
\end{aligned}
\tag{4.5}
$$

其中 $B_{p,s}$ 是习题 4.3 中定义的一些常数。从式(4.5)容易看出,如果 $\lambda_s = o_p(n^{-1/2})$,那么我们有 $\sqrt{n}(\hat{p}(x^d) - p(x^d)) \xrightarrow{d} N(0, p(x^d)(1 - p(x^d)))$,即,如果 $\lambda = o_p(n^{-1/2})$,那么 $\hat{p}(x^d)$ 有和在第 3 章式(3.2)中定义的频率估计量 $\tilde{p}(x^d)$ 相同的渐近分布。

当处理连续变量时情形也是如此,平滑参数的选择是极为重要的。我们建议通过最小化一个基于 $\hat{p}(\cdot)$ 和 $p(\cdot)$ 之间差的平方的判别函数来选择平滑参数 $\lambda_1, \cdots, \lambda_r$,它由下式给出 $\left(\sum_{x^d} \equiv \sum_{x^d \in S^d}\right)$

$$
\begin{aligned}
I_n &= \sum_{x^d} [\hat{p}(x^d) - p(x^d)]^2 \\
&= \sum_{x^d} [\hat{p}(x^d)]^2 - 2\sum_{x^d} \hat{p}(x^d)p(x^d) + \sum_{x^d} [p(x^d)]^2 \\
&\equiv I_{1n} - 2I_{2n} + \sum_{x^d} [p(x^d)]^2
\end{aligned}
\tag{4.6}
$$

其中 $I_{1n} = \sum_{x^d} [\hat{p}(x^d)]^2$,$I_{2n} = \sum_{x^d} \hat{p}(x^d)p(x^d)$,注意 $I_{2n} = E_X[\hat{p}(X^d)]$,这里 $E_X[\cdot]$ 表示关于 X^d 的期望,而不是用于定义 $\hat{p}(\cdot)$ 中所使用的随机观测值 $\{X_j^d\}_{j=1}^n$,因此,我们可以通过用样本均值代替总体均值 $E_X[\cdot]$ 来估计 I_{2n}:

$$
\hat{I}_{2n} = \frac{1}{n}\sum_{i=1}^{n} \hat{p}_{-i}(X_i^d) = (n(n-1))^{-1}\sum_{i=1}^{n}\sum_{j \neq i}^{n} L_{\lambda,ij} \tag{4.7}
$$

其中 $L_{\lambda,ij}=L(X_i^d,X_j^d,\lambda)$，$\hat{p}_{-i}(X_i^d)=n^{-1}\sum_{j=1,j\neq i}^{n}L_{\lambda,ij}$ 是 $p(X_i^d)$ 的去一核估计量。

129 式(4.6)右边最后一项是与 $\lambda_1,\cdots,\lambda_r$ 无关的，因此我们可以选择 $\lambda_1,\cdots,\lambda_r$ 来最小化：

$$\mathrm{CV}_p(\lambda)=I_{1n}-2\hat{I}_{2n}=\sum_{x^d\in S^d}[\hat{p}(x^d)]^2-2(n(n-1))^{-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n}L_{\lambda,ij} \quad (4.8)$$

令 $\hat{\lambda}_1,\cdots,\hat{\lambda}_r$ 表示最小化式(4.8)的交错鉴定法的 λ_s。Ouyang 等(2006)证明了只要 x_s^d 不是均匀分布，那么 $\hat{\lambda}_s=o_p(1)\ (s=1,\cdots,r)$。令 $\lambda_{(r)}=(\lambda_1,\cdots,\lambda_r)'$ 表示平滑参数的 $r\times 1$ 向量。为了推导出 $\hat{\lambda}_s$ 的极限性质，Ouyang 等使用了 U 统计量 H 分解，并把 $\mathrm{CV}_p(\lambda)$ 扩展为一个 λ_s 的多项式以得到如下结果：

$$\mathrm{CV}_p(\lambda)=\lambda_{(r)}'\Omega_r\lambda_{(r)}-n^{-1}A_i'\lambda(r)+o_p\left(n^{-1}\sum_{s=1}^{r}\lambda_s+\sum_{s=1}^{r}\lambda_s^2\right)$$
$$+\text{与 }\lambda\text{ 无关的项} \quad (4.9)$$

其中 Ω_r 是一个 $r\times r$ 阶的半正定矩阵，A_r 是一个 $r\times 1$ 阶的 $O_p(1)$ 随机向量。Ω_r 的显性表达式在习题4.4中给出。如果 Ω_r 是正定的，那么，通过关于 $\lambda_{(r)}$ 最小化式(4.9)，我们得到：

$$\hat{\lambda}_{(r)}=n^{-1}\Omega_r^{-1}A_r+o_p(n^{-1}) \quad (4.10)$$

方程(4.10)说明了对 $s=1,\cdots,r$ 有 $\hat{\lambda}_s=O_p(n^{-1})$，式(4.10)成立所需要的关键条件是 Ω_r 是正定的。下面我们证明当且仅当 X^d 对所有的 $s=1,\cdots,r$ 关于 x_s 不是均匀分布时，Ω_r 是正定的。

Ouyang 等(2006)证明了：

$$B_\lambda \overset{\text{def}}{=} \lambda_{(r)}'\Omega_r\lambda_{(r)}=\sum_{x^d\in S^d}\left\{\sum_{s=1}^{r}\lambda_s[p(x^d)-p_{1,s}(x^d)]\right\}^2 \quad (4.11)$$

其中 $P_{1,s}$ 的定义见习题4.4。

方程(4.11)证明了 Ω_r 是半正定的，因为对所有的 $\lambda_{(r)}$ 有 $\lambda_{(r)}'\Omega_r\lambda_{(r)}\geqslant 0$。可以证明，如果 X^d 关于它的任何元素，记为 x_s^d，是均匀分布的，那么 $p_{1,s}(x^d)=p(x^d)$，所以 $B(\lambda)$ 不依赖 λ_s（λ_s 退出是因为在 $B(\lambda)$ 中它的系数为0）。因此，

130 在这种情况下 Ω_r 不是正定的。Ouyang 等(2006)更进一步证明逆命题为真，即如果对所有的 x_s^d，X^d 不是均匀分布的，则 Ω_r 是正定的。

我们用 $\hat{\lambda}_1,\cdots,\hat{\lambda}_r$ 表示最小化式(4.8)的 $\lambda_1,\cdots,\lambda_r$ 的交错鉴定法的选择，从式(4.10)立即得到下一个结论。

定理 4.1 假定 X^d 对所有的 $s=1,\cdots,r$ 关于 x_s 不是均匀分布的，那么

$$\hat{\lambda}_s=O_p(n^{-1}),\ s=1,\cdots,r$$

证明：定理 4.1 的条件表明 Ω_r 是正定矩阵。因此由式(4.10)推出定理 4.1。由定理 4.1 中给出 $\hat{\lambda}_s$ 的收敛速度为 n^{-1}，从式(4.5)立即得出以下的结果。 □

定理 4.2　在与定理 4.1 一样的条件下，我们有

(i) $\hat{p}(x^d)-p(x^d)=O_p(n^{-1/2})$。

(ii) $n^{1/2}(\hat{p}(x^d)-p(x^d))\xrightarrow{d}N(0,p(x^d)(1-p(x^d)))$。

定理 4.2 的证明遵循定理 4.1，它被留作问题(见习题 4.6)。

定理 4.2 说明 $\hat{p}(x^d)$ 的渐近分布与频率估计量 $\bar{p}(x^d)$ 相同。这确实是人们所期望的，由于渐近上我们有一个无限大的样本容量，因此即使样本分割，每个单位仍然有无限多的数据，同样也是渐近的。所以，如果样本容量足够大，$\hat{\lambda}_1,\cdots,\hat{\lambda}_r$ 应该都接近于 0，并且我们的交错鉴定估计量 $\hat{p}(x^d)$ 将非常接近于频率估计量 $\bar{p}(x^d)$。然而在有限样本容量的应用中，这两种方法会产生差别很大的结果。

定理 4.1 表明交错鉴定的 $\hat{\lambda}_s$ 以 n^{-1} 的速度收敛于 0，它与 λ_s 的最大似然交错鉴定选择有相同的收敛速度(见 Hall(1981))。

定理 4.1 没有包括一个重要的情形，即当对某些 $s\in\{1,\cdots,r\}$，X^d 关于 x_s 是均匀分布时的情形。不失一般性，假定对 $s=1,\cdots,r_1$，X^d 关于 x_s^d 不是均匀分布的，并且对 $s=r_1+1,\cdots,r(1\leqslant r_1<r)$，$X^d$ 关于 x_s^d 是均匀分布的。下面的定理证明了 $\hat{\lambda}_s$ 的渐近性质依 x_s^d 是否均匀分布而不同。

定理 4.3　假定 $p(x^d)$ 关于 $x_1,\cdots,x_{r_1}$ 不是一个均匀分布，而关于 $x_{r_1+1},\cdots,x_r$ 是一个均匀分布，那么　131

(i) $\hat{\lambda}_s=O_p(n^{-1}),s=1,\cdots,r_1$。

(ii) 对某些 $0<\delta<1$，有 $\lim\limits_{n\to\infty}P\left[\hat{\lambda}_s=\frac{c_s-1}{c_s}\right]\geqslant\delta,s=r_1+1,\cdots,r$。

定理 4.3 在 Ouyang 等(2006)中被证明。定理 4.3 表明与均匀分布的变量相关的平滑参数将不会收敛于 0，然而，它们有很大可能性会取其极大值，以至于估计的概率函数对 $s=r_1+1,\cdots,r$ 满足式(4.4)的均匀分布条件，并且比没有施加这些限制的估计量更加有效。确定 δ 的精确值是困难的。Ouyang 等报告的模拟结果建议，对大范围数据生成过程 δ 取值在 0.6 附近。

4.2　离散数据的平滑回归

我们现在考虑一个下式给出的非参数的回归模型：

$$Y_i=g(X_i^d)+u_i$$

其中 $g(\cdot)$ 是一个未知函数，$X_i^d \in \mathcal{S}^d$，$E(u_i \mid X_i^d)=0$ 且 $E(u_i^2 \mid X_i^d)=\sigma^2(X_i^d)$ 形式未知。尽管式(4.2)中定义的核函数可以用来估计 $g(x^d)$，但是我们建议在回归设定中使用如下另外一种简单的核函数。

$$l(X_{is}^d, x_s^d, \lambda_s) = \begin{cases} 1, & X_{is}^d = x_s^d \\ \lambda_s, & \text{其他} \end{cases} \tag{4.12}$$

当 $\lambda_s=0$，$l(X_{is}^d, x_s^d, 0)$ 成为一个示性函数；当 $\lambda_s=1$，$l(X_{is}^d, x_s^d, 1)\equiv 1$ 是一个均匀权函数。因此，λ_s 的值域为[0,1]。注意，式(4.12)定义的核权函数总和不等于1，因此，它不适用于估计一个概率函数。然而，它不会影响下面式(4.14)中定义的非参数估计量 $\hat{g}(x)$，因为核函数同时出现在式(4.14)的分子和分母中，并且显而易见地，核函数因此可以乘以任何非零常数使 $\hat{g}(x)$ 的定义不受影响。

132

通过使用式(4.12)，可以得到乘积核函数：

$$L(X_i^d, x^d, \lambda) = \prod_{s=1}^{r} \lambda_s^{N_{is}(x)} \tag{4.13}$$

其中 $N_{is}(x)=1(X_{is}^d \neq x_s^d)$（示性函数）。我们通过下式估计 $g(x^d)$：

$$\hat{g}(x^d) = \frac{n^{-1}\sum_{i=1}^{n} Y_i L(X_i^d, x^d, \lambda)}{\hat{p}(x^d)} \tag{4.14}$$

其中 $\hat{p}(x^d) = n^{-1}\sum_{i=1}^{n} L(X_i^d, x^d, \lambda)$，当对所有的 $s=1,\cdots,r$，$\lambda_s=0$ 时，我们的估计量退化为在式(3.6)中给出的频率估计量。

在当 $n\to\infty$ 时，$\lambda_s\to 0\,(s=1,\cdots,r)$ 的假定下，容易证明 $\hat{g}(x^d)-g(x^d)=O_p\left(\sum_{s=1}^{r}\lambda_s+n^{-1/2}\right)$。例如，正如在第2章中讨论的对单纯的连续变量估计量证明的情形，我们可以记 $\hat{g}(x^d)-g(x^d)=[\hat{g}(x^d)-g(x^d)]\hat{p}(x^d)/\hat{p}(x^d)\equiv\hat{m}(x^d)/\hat{p}(x^d)$。可以证明 $E[\hat{m}(x^d)]=O\left(\sum_{s=1}^{r}\lambda_s\right)$ 和 $\mathrm{var}[\hat{m}(x^d)]=O(n^{-1})$。这意味着 $E[\hat{m}(x^d)^2]=O\left(\sum_{s=1}^{r}\lambda_s^2+n^{-1}\right)$，这反过来意味着 $\hat{m}(x^d)=O_p\left(\sum_{s=1}^{r}\lambda_s+n^{-1/2}\right)$，同时通过与导出式(4.5)相同的论证可以很容易证明 $\hat{p}(x^d)=p(x^d)+O_p\left(\sum_{s=1}^{r}\lambda_s+n^{-1/2}\right)$。因此，我们有：

$$\hat{g}(x^d)-g(x^d)=\frac{\hat{m}(x^d)}{\hat{p}(x^d)}=\frac{O_p\left(\sum_{s=1}^{r}\lambda_s+n^{-1/2}\right)}{p(x^d)+o_p(1)}$$

$$= O_p\left(\sum_{s=1}^{r}\lambda_s + n^{-1/2}\right) \tag{4.15}$$

现在我们讨论使用交错鉴定法来选择 λ_s。我们选择 $\lambda_1,\cdots,\lambda_r$ 来最小化交错鉴定法中的残差平方和：

$$\mathrm{CV}_d(\lambda) = n^{-1}\sum_{i=1}^{n}[Y_i - \hat{g}_{-i}(X_i^d)]^2 \tag{4.16}$$

其中

$$\hat{g}_{-i}(X_i^d) = \frac{n^{-1}\sum_{j\neq i}^{n} Y_j L(X_i^d, X_j^d, \lambda)}{\hat{p}_{-i}(X_i^d)} \tag{4.17}$$

是 $g(X_i^d)$ 的去一核估计量，且 $\hat{p}_{-i}(X_i^d)$ 是 $p(X_i^d)$ 的去一估计量。我们用 $\hat{\lambda}_1,\cdots,\hat{\lambda}_r$ 表示最小化式(4.16)的交错鉴定法选择的 $\lambda_1,\cdots,\lambda_r$。

133

在本节我们只考虑 X^d 的所有元素都是相关的情形，我们把潜在的不相关回归元情形推迟至下一节讨论。同样，在4.5节我们将讨论有分类和连续回归元混合的更一般的情形。我们从以下的假定开始。

假定 4.1

(i) $E(Y_i^2 | X_i^d)$ 是有界的。

(ii) 有唯一的 $(\lambda_1,\cdots,\lambda_r)$ 值满足：

$$\sum_{x^d\in S^d} p(x^d)\left\{\sum_{z^d\in S^d} p(z^d)[g(x^d) - g(z^d)]L(x^d, z^d, \lambda)\right\}^2 = 0$$

那就是对所有的 $s = 1,\cdots,r, \lambda_s = 0$。

假定4.1(ii)意味着 $g(x)$ 对任意元素 $x_s \in D_s$ 不是一个常数函数(见习题4.11)，并且它也意味着对所有的 $s = 1,\cdots,r, \hat{\lambda}_s = o_p(1)$。

下一个定理建立了 $\hat{\lambda}_1,\cdots,\hat{\lambda}_r$ 的收敛速度。

定理 4.4 基于假定4.1，我们有

$$\hat{\lambda}_s = O_p(n^{-1}),\quad 所有\ s = 1,\cdots r$$

Li, Ouyang 和 Racine(2006)证明了定理4.4。习题4.7要求读者在 x^d 是单变量的简单情形下证明定理4.4(也提供了提示)。

定理4.4证明了 $\hat{\lambda}_s$ 以 $O_p(n^{-1})$ 速度收敛于0。在这么快的收敛速率下很容易得到 $\hat{g}(x)$ 的渐近分布，下一个定理将给出证明。

定理 4.5 基于假定4.1，我们有

$$\sqrt{n}(\hat{g}(x^d) - g(x^d))/\sqrt{\hat{\sigma}^2(x^d)/\hat{p}(x^d)} \to N(0,1)\ 依分布收敛$$

其中 $\hat{\sigma}^2(x^d) = n^{-1}\sum_i [Y_i - \hat{g}(X_i^d)]^2 L(X_i^d, x^d, \hat{\lambda})/\hat{p}(x^d)$ 是 $\sigma^2(x^d) = E[u_i^2 | X_i^d = x^d]$

的一致估计量。

134　证明：回忆频率估计量 $\tilde{g}(x^d)$ 可以从 $\hat{g}(x^d)$ 得到，对所有 $s=1,\cdots,r$ 有 $\lambda_s=0$。使用 $\hat{\lambda}_s=O_p(n^{-1})$，容易看出 $\hat{g}(x)=\tilde{g}(x)+O_p(n^{-1})$。因此通过式(3.8)可得 $\sqrt{n}(\hat{g}(x)-g(x))=\sqrt{n}(\tilde{g}(x)-g(x))+O_p(n^{-1/2})\xrightarrow{d}N(0,\sigma^2(x^d)/p(x^d))$。最后，习题 4.8 证明 $\hat{\sigma}^2(x^d)=\sigma^2(x^d)+o_p(1)$。定理 4.5 可以直接得出。 □

在下一节我们将讨论一个回归元子集不相关的情形。

4.3　有离散回归元的核回归：无关回归元的情形

在这一节我们考虑这样的可能性，一些回归元在独立于因变量的意义上，它们实际上是不相关的。不失一般性，我们假定 X_i^d 的前 $r_1(1\leqslant r_1<r)$ 个元素是相关的，而 X_i^d 剩余的 $r_2=r-r_1$ 个元素是不相关的。令 $\bar{X}_i^d$ 表示 r_1 维的 X_i^d 的相关元素，$\tilde{X}_i^d$ 表示 r_2 维的不相关元素。类似 Li 等(2006)中所采用的方法，我们假定

$$(Y,\bar{X}^d)\text{ 与 }\tilde{X}^d\text{是相互独立的} \tag{4.18}$$

$g(x^d)$ 的核估计量和 CV 目标函数的定义如同 4.2 节所给出的那样是相同的。我们仍然用 $(\hat{\lambda}_1,\cdots,\hat{\lambda}_r)$ 表示 CV 所选择的平滑参数。在下面的定理 4.6 中我们证明了(i) 与相关回归元相关联的平滑参数以 $n^{-1/2}$ 的速度收敛于 0，这与当不相关元素不存在时 n^{-1} 的收敛速度是不同的，(ii) 与不相关元素相关联的平滑参数不会收敛于 0，但是有很大的可能性收敛于它们的极大值，所以这些不相关的回归元有很大的可能性被平滑掉了。此外，即使当 $n\to\infty$ 时，这些平滑参数也有可能不收敛于它们的极大值。

与用于 $\bar{x}^d$ 和 $\tilde{x}^d$ 的符号类似，我们将用 $\bar{\lambda}$ 表示与相关回归元相关联的平滑参数，用 $\tilde{\lambda}$ 表示与不相关回归元相关联的平滑参数。因此我们有 $\bar{\lambda}_s=\lambda_s(s=$
135　$1,\cdots,r_1)$ 和 $\tilde{\lambda}_s=\lambda_{r_1+s}(s=1,\cdots,r_2(r_2=r-r_1))$。同样，用 $\bar{p}(\cdot)$ 和 $\tilde{p}(\cdot)$ 分别表示 $\bar{X}$ 和 $\tilde{X}$ 的边缘概率密度函数。那么根据式(4.18)我们知道 $p(x^d)=\bar{p}(\bar{x}^d)\tilde{p}(\tilde{x}^d)$，用 $\bar{S}^d$ 表示 $\bar{x}^d$ 的支撑，我们对假定 4.1 做出如下修改

假定 4.2　有唯一的 $(\lambda_1,\cdots,\lambda_{r_1})$ 值满足

$$\sum_{\bar{x}^d\in\bar{S}^d}\bar{p}(\bar{x}^d)\left\{\sum_{\bar{z}^d\in\bar{D}}\bar{p}(\bar{z}^d)[g(\bar{x}^d)-g(\bar{z}^d)]L(\bar{x}^d,\bar{z}^d,\bar{\lambda})\right\}^2=0$$

那就是对所有的 $s=1,\cdots,r_1,\lambda_s=0$。

假定 4.2 保证与相关回归元相联系的交错鉴定法所选择的平滑参数将收敛于 0，然而，我们没有对与不相关回归元相联系的平滑参数施加任何假定，除了它们的取值在[0,1]区间内。

下面的定理给出了交错鉴定法所选择的平滑参数的渐近性质。

定理 4.6　假定 $r_1 \geqslant 1, r_2 \geqslant 1(r = r_1 + r_2 \geqslant 2)$。那么，在假定 4.2 下，我们有

$$\hat{\lambda}_s = O_p(n^{-1/2}), s = 1, \cdots, r_1$$

$$\lim_{n\to\infty} P(\hat{\lambda}_{r_1+1} = 1, \cdots, \hat{\lambda}_r = 1) \geqslant \alpha, \text{某些 } \alpha \in (0,1)$$

定理 4.6 的证明见 Li 等(2006)。

定理 4.6 表明与相关回归元相关联的平滑参数都以 n^{-1} 的速度收敛于 0，然而与不相关回归元相关联的平滑参数有可能取它们的最大值 1。也就是说，不相关回归元将有可能被平滑掉。在一般情况下确定 α 的准确值是困难的。Li 等(2006)报告的模拟结果表明通常有 60% 的可能性 $\hat{\lambda}_s$ 取极大值 1，有 40% 的可能性 $\hat{\lambda}_s$ 取值在 0 和 1 之间，$s = r_1 + 1, \cdots, r$。

注意当 x_s^d 是不相关回归元时，$\hat{\lambda}_s$ 的渐近性质很难进一步详细描述，因为在这种情况下 $\hat{\lambda}_s$ 并不收敛于 0。因此产生的 $\hat{g}(x^d)$ 的渐近分布也很难获得。当人们得到 $\hat{\lambda}_s$ 的一个相对较大的值并推测这可能反映了这样的事实，即 x_s^d 是不相关回归元时，人们可能对原假定做一个正式的检验，也就是 x_s^d 确实是一个不相关回归元，例如，使用基于 Racine 等(见 12 章)建议的自举法进行检验。如人们未能拒绝原假定，就可以把该回归元从模型中去除。用这种方法，只有相关回归元才有可能继续留在模型中，这样就可以用那些被认为是相关的回归元去重新计算交错鉴定窗宽。那么由于没有不相关回归元保留在这个模型中，就可以应用定理 4.4 和定理 4.5 的结论。

136

我们现在进入到一般的情形，有离散和连续数据混合的平滑参数的非参数估计。

4.4　混合数据的回归：相关回归元的情形

4.4.1　混合数据的平滑估计

这一节我们考虑这样一个非参数回归模型，它的一个回归元子集是分类的，剩余的为连续的。同第 3 章一样，我们使用 X_i^d 来表示一个 $r \times 1$ 阶取离散数值的回归元向量，用 $X_i^c \in \mathbb{R}^q$ 表示剩余的连续回归元。我们再次用 X_{is}^d 表示 X_i^d 的第 s 个元素。我们假定 X_{is}^d 取 $c_s \geqslant 2$ 个不同的值，即 $X_{is}^d \in \{0, 1, \cdots, c_s - 1\}$，$s = 1, \cdots, q$，并定义 $X_i = (X_i^d, X_i^c)$。

这个非参数回归模型为：

$$Y_i = g(X_i) + u_i \tag{4.19}$$

满足 $E(u_i | X_i) = 0$。我们用 $f(x) = f(x^c, x^d)$ 来表示 (X_i^d, X_i^c) 的联合概率密度函数。

对于离散变量 X_i^d，我们首先考虑这些变量没有自然排序的情形。扩展到一般情形时一些离散回归元具有自然排序的情况将在本节的末尾讨论。

对 $x^c = (x_1^c, \cdots, x_q^c)$，定义

$$W_h(x^c, X_i^c) = \frac{1}{h_1 \cdots h_q} W\left(\frac{x^c - X_i^c}{h}\right) = \prod_{s=1}^{r} \frac{1}{h_s} w\left(\frac{x_s^c - X_{is}^c}{h_s}\right)$$

其中 w 是对称的非负的一元核函数。$0 < h_s < \infty$ 是 x_s^c 的平滑参数。对 $x^d =$ 137 $(x_1^d, \cdots, x_r^d)$，定义

$$L(x^d, X_i^d, \lambda) = \prod_{s=1}^{r} \lambda_s^{N_{is}(x)}$$

其中 $N_{is}(x) = \mathbf{1}(X_{is}^d \neq x_s^d)$ 是示性函数，当 $X_{is}^d \neq x_s^d$ 时等于 1，否则为 0。x_s^d 的平滑参数为 $0 \leq \lambda_s \leq 1$。注意当 $\lambda_s = 1$ 时，$L(x^d, X_i^d, 1)$ 将与 (x_s^d, X_{is}^d) 不相关（即 x_s^d 被平滑掉了）。

混合变量 $x = (x^c, x^d)$ 的向量的核函数就是 $W_h(\cdot)$ 和 $L(\cdot)$ 的乘积，即定义 $K_\gamma(x, X_i) = W_h(x^c, X_i^c) L(x^d, X_i^d, \lambda)$ $(\gamma = (h, \lambda))$，我们可以这样估计 $g(x)$：

$$\hat{g}(x) = \frac{n^{-1} \sum_{i=1}^{n} Y_i K_\gamma(x, X_i)}{n^{-1} \sum_{i=1}^{n} K_\gamma(x, X_i)} \tag{4.20}$$

注意分母 $n^{-1} \sum_{i=1}^{n} K_\gamma(x, X_i)$ 是联合密度函数 $f(x) = f(x^c, x^d)$ 的核估计量。可以证明 $\hat{g}(x) - g(x) = O_p\left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r} \lambda_s + (nh_1 \cdots h_q)^{-1/2}\right)$。定义 $\hat{m}(x) = [\hat{g}(x) - g(x)]\hat{f}(x)$，然后可以证明（见习题 4.9）：

$$E[\hat{m}(x)] = \sum_{s=1}^{q} B_{1s}(x) f(x) h_s^2 + \sum_{s=1}^{r} B_{2s}(x) f(x) \lambda_s + o(\eta_2) \tag{4.21}$$

$$\operatorname{var}(\hat{m}(x)) = \frac{\kappa^q \sigma^2(x) f(x)}{nh_1 \cdots h_q} [1 + O(\eta_2)] \tag{4.22}$$

$$\hat{f}(x) = f(x) + O_p(\eta_2 + \eta_1^{1/2}) \tag{4.23}$$

其中

$$B_{1s}(x) = \frac{\kappa^2}{2} [g_{ss}(x) + 2 g_s(x) f_s(x) f(x)^{-1}]$$

$$B_{2s}(x) = \sum_{z^d \in S^d} \mathbf{1}_s(x^d, z^d) [g(x^c, z^d) - g(x)] f(x^c, z^d) f(x)^{-1}$$

$$\kappa = \int \omega(v)^2 \mathrm{d}v, \quad \eta_2 = \sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{r} \lambda_s$$

$$\eta_1 = (nh_1 \cdots h_q)^{-1}$$

且

$$\mathbf{1}_s(x^d, z^d) \stackrel{\text{def}}{=} \mathbf{1}(x_s^d \neq z_s^d) \prod_{s' \neq s}^{r} \mathbf{1}(x_{s'}^d = z_{s'}^d) \tag{4.24}$$

$\mathbf{1}_s(x^d, z^d)$当 x^d 和 z^d 仅在它们的第 s 个元素不同时取值为 1,否则取 0。

138

式(4.21)和式(4.22)等同于 $E(\hat{m}(x)^2) = O(\eta_2^2 + \eta_1)$,这意味着 $\hat{m}(x) = O_p(\eta_2 + \eta_1^{1/2})$,以及

$$\hat{g}(x) - g(x) = \frac{\hat{m}(x)}{\hat{f}(x)} = \frac{O_p(\eta_2 + \eta_1^{1/2})}{f(x) + o_p(1)} = O_p(\eta_2 + \eta_1^{1/2})$$

如果除 $\eta_1 = o(1)$,$\eta_2 = o(1)$外,人们也可假定$(nh_1 \cdots h_q)^{1/2} \eta_2^{3/2} = o(1)$,那么使用式(4.21)到式(4.23)根据 Liapunov 中心极限定理可以得到 $\hat{g}(x)$的渐近正态分布:

$$\sqrt{nh_1 \cdots h_q}\left[\hat{g}(x) - g(x) - \sum_{s=1}^{q} B_{1s}(x) h_s^2 - \sum_{s=1}^{r} B_{2s}(x) \lambda_s\right]$$

$$\xrightarrow{d} N(0, \kappa^q \sigma^2(x)/f(x)) \tag{4.25}$$

4.4.2 交错鉴定法

最小二乘交错鉴定法通过选择 $h_1, \cdots, h_q, \lambda_1, \cdots, \lambda_r$ 来最小化交错鉴定函数:

$$\mathrm{CV}_r(h, \lambda) = \sum_{i=1}^{n} (Y_i - \hat{g}_{-i}(X_i))^2 M(X_i) \tag{4.26}$$

其中 $\hat{g}_{-i}(X_i) = \sum_{l \neq i}^{n} Y_l K_\gamma(X_i, X_l) \Big/ \sum_{l \neq i}^{n} K_\gamma(X_i, X_l)$ 是 $g(X_i)$ 的去一核估计量,$0 \leqslant M(\cdot) \leqslant 1$ 是权重函数,用于避免除以 0 和由于边界效应导致收敛速度慢带来的困难。

我们假定:

$$(h_1, \cdots, h_q, \lambda_1, \cdots, \lambda_r) \in [0, \eta]^{q+r}, \; nh_1 \cdots h_q \geqslant t_n \tag{4.27}$$

其中 $\eta = \eta_n$ 是一个正的收敛于 0 的序列,收敛速度慢于任意 n 的多项式的倒数,t_n 是发散到无穷的常数序列。

用$\mathcal{S}$表示 $M(\cdot)$的支撑,我们也假定:

$g(\cdot)$ 和 $f(\cdot)$ 是两个连续可微的函数;$m(\cdot)$ 是连续、非负且有紧支撑的函数;对 $x = (x^c, x^d) \in \mathcal{S}$,$f(\cdot)$ 非负且有界　　(4.28)

139 可以证明 CV_r 的首项是 CV_{r0}，由下式给出（忽略与 (h,λ) 不相关的项）：

$$CV_{r0}(h,\lambda) = \sum_{x^d}\int\left\{\left[\sum_{s=1}^{q}B_{1s}(x)h_s^2 + \sum_{s=1}^{r}B_{2s}(x)\lambda_s\right]^2 f(x) + \frac{\kappa^q\sigma^2(x)}{nh_1\cdots h_q}\right\}M(x)\,dx^c \tag{4.29}$$

利用交错鉴定和 $\hat{g}(x)$ 的均方误差之间的关系也可以推导出式(4.29)。在第2章中我们指出相对于逐点的均方误差，CV 的首项是：

$$CV \sim \sum_{x^d}\int MSE[\hat{g}(x)]f(x)M(x)\,dx^c \tag{4.30}$$

方程(4.25)意味着：

$$MSE[\hat{g}(x)] \sim \left[\sum_{s=1}^{q}B_{1s}(x)h_s^2 + \sum_{s=1}^{r}B_{2s}(x)\lambda_s\right]^2 + \kappa^q\sigma^2(x)f(x)^{-1}(nh_1\cdots h_q)^{-1}$$

将上式代入式(4.30)，忽略与 (h,λ) 不相关的项和低阶项，我们得到 $CV_r(a,b) = n^{-4/(4+q)}\chi_r$，其中

$$\chi_r(a,b) = \sum_{x^d}\int\left\{\left[\sum_{s=1}^{q}B_{1s}(x)a_s^2 + \sum_{s=1}^{r}B_{2s}(x)b_s\right]^2 f(x) + \frac{\kappa^q\sigma^2(x)}{a_1\cdots a_q}\right\}M(x)\,dx^c \tag{4.31}$$

其中 $a_s = n^{1/(4+q)}h_s$，$b_s = n^{2/(4+q)}\lambda_s$。方程(4.31)确实给出了 CV_r 的首项，正如式(4.29)说明的。

令 $a_1^0,\cdots,a_q^0,b_1^0,\cdots,b_r^0$ 表示在每一项非负的约束下最小化 χ_r 的值。我们排除了某些 a_s^0 或 b_s^0 是无限的情形，并要求：

$$a_s^0 \text{ 和 } b_s^0 \text{ 是唯一定义的，且均为有界的} \tag{4.32}$$

这意味着对所有的 s，有 $0 < a_s^0 < \infty$，但是对某些 s，b_s^0 可能会为 0。习题 4.10 在 $q = r = 1$ 的简单情形下，对 a^0 和 b^0 推导出了显性表达式。

140 令 $h_s^0(1\leqslant s\leqslant q)$ 和 $\lambda_s^0(1\leqslant s\leqslant r)$ 表示最小化 $CV_{r0}(h,\lambda)$ 的 h_s 和 λ_s 的值，那么式(4.32)意味着下式：

$$\begin{aligned} h_s^0 &\sim n^{-1/(q+4)}a_s^0 = O(n^{-1/(q+4)}), \quad 1\leqslant s\leqslant q \\ \lambda_s^0 &\sim n^{-2/(q+4)}b_s^0 = O(n^{-2/(q+4)}), \quad 1\leqslant s\leqslant r \end{aligned} \tag{4.33}$$

令 $\hat{h}_s(s=1,\cdots,q)$ 和 $\hat{\lambda}_s(s=1,\cdots,r)$ 分别表示通过交错鉴定法确定的 h_s 和 λ_s 的值，下面的定理证明了当 h_s^0 和 λ_s^0 分别被 $\hat{h}_s$ 和 $\hat{\lambda}_s$ 代替时式(4.33)成立。

定理 4.7 在条件(4.27)、(4.28)和(4.32)下，有

$$n^{1/(q+4)}\hat{h}_s \to a_s^0, 1\leqslant s\leqslant q$$

$$n^{2/(q+4)}\hat{\lambda}_s \to b_s^0, 1\leqslant s\leqslant r$$

以及 $\inf n^{4/(q+4)}\left[\mathrm{CV}_r(h,r) - n^{-1}\sum_{i=1}^{n} u_i^2 M_i\right] \to \inf_{\chi}$ 依概率收敛。

定理4.7 的证明可以在 Hall 等(2006)中找到,定理4.7 来自式(4.33)和这样一个事实:

$$\mathrm{CV}_r(h,\lambda) - \int \sigma^2(x) M(x)\,\mathrm{d}x = \mathrm{CV}_{r0}(h,\lambda) + (\mathrm{s.o.})$$

对所有的$(x^c,x^d) \in w \times \mathcal{S}^d$, $(h,\lambda) \in [0,\eta_n]^{q+r}$一致收敛,其中 $\eta_n > 0$,且 $n \to \infty$ 时 $\eta_n \to 0$。

直觉上,人们将期望最小化 $\mathrm{CV}_r(\cdot,\cdot)$ 的 $\hat{h}_s$ 和 $\hat{\lambda}_s$ 和最小化 $\mathrm{CV}_r(\hat{h},\hat{\lambda})$ 首项 $\mathrm{CV}_{r0}(\cdot,\cdot)$ 的 h_s^0 和 λ_s^0 有相同渐近性质。因此,我们期望$\hat{h}_s = h_s^0 + (\mathrm{s.o.})$和 $\hat{\lambda}_s = \lambda_s^0 + (\mathrm{s.o.})$。

4.5　混合数据的回归:无关回归元的情形

在这一节,我们将考虑某些回归元可能是不相关的情形(即允许这种可能性)。不失一般性,我们假定只有 X^c 的前 q_1 个元素和 X^d 的前个元素 r_1 在下面定义的意义上是相关回归元。对整数 $0 \leqslant q_1, q_2 \leqslant q$ 和 $0 \leqslant r_1, r_2 \leqslant r$ 满足 $q_1 + q_2 = q$ 以及 $r_1 + r_2 = r$,令 $\bar{X}$ 包含 X^c 的前 q_1 个元素和 X^d 的前 r_1 个元素,并令 $\tilde{X} = X/\bar{X}$ 表示 X 中剩余的元素,在这一节我们假定:　141

$$(Y,\bar{X})\ 独立于\ \tilde{X} \tag{4.34}$$

条件(4.34)的一个结果是 $E[Y|\bar{X},\tilde{X}] = E[Y|\bar{X}]$,使得在式(4.19)中只有 $\bar{X}$ 是相关的(因此,$\tilde{X}$ 只包含无关回归元)。然而,我们并没有假定这个信息是已知的,并且在实践中,我们因此用非参数估计 $E[Y_i|X_i]$而不是 $E[Y_i|\bar{X}]$。我们将证明,如果人们一起平滑离散和连续变量,并用最小二乘交错鉴定法来选择平滑参数,那么渐近地无关变量可自动被平滑掉。为了更明确地说明我们所说的"无关变量将被平滑掉",注意:

$$\hat{g}(x) = \frac{\sum_{i=1}^{n} Y_i \prod_{s=1}^{q_1} w\left(\frac{X_{is}-x_s}{h_s}\right) \prod_{s=q_1+1}^{q} w\left(\frac{X_{is}-x_s}{h_s}\right)}{\sum_{i=1}^{n} \prod_{s=1}^{q_1} w\left(\frac{X_{is}-x_s}{h_s}\right) \prod_{s=q_1+1}^{q} w\left(\frac{X_{is}-x_s}{h_s}\right)} \times \frac{\prod_{s=1}^{r_1} l(X_{is}^d, x_s^d, \lambda_s) \prod_{s=r_1+1}^{r} l(X_{is}^d, x_s^d, \lambda_s)}{\prod_{s=1}^{r_1} l(X_{is}^d, x_s^d, \lambda_s) \prod_{s=r_1+1}^{r} l(X_{is}^d, x_s^d, \lambda_s)} \tag{4.35}$$

如果对 $s=q_1+1,\cdots,q,h_s=\infty$，对 $s=r_1+1,\cdots,r,\lambda_s=1$，那么对 $s=q_1+1,\cdots,q$，有 $w((X_{is}-x_s)/h_s)=w(0)$，对 $s=r_1+1,\cdots,r$，有 $l(X_{is}^d,x_s^d,1)=1$，并且式(4.35)可简化为：

$$\hat{g}(x)=\frac{\sum_{i=1}^{n} Y_i \prod_{s=1}^{q_1} w\left(\frac{X_{is}-x_s}{h_s}\right)\prod_{s=1}^{r_1} l(X_{is}^d,x_s^d,\lambda_s)}{\sum_{i=1}^{n}\prod_{s=1}^{q_1} w\left(\frac{X_{is}-x_s}{h_s}\right)\prod_{s=1}^{r_1} l(X_{is}^d,x_s^d,\lambda_s)} \tag{4.36}$$

由于 $w(0)^{q-q_1}$ 在 $\hat{g}(x)$ 的分子和分母中被消掉，结果是 $\hat{g}(x)$ 不再与无关变量 x_s^c $(s=q_1+1,\cdots,q)$ 和 x_s^d $(s=r_1+1,\cdots,r)$ 相关。下面的定理 4.8 中我们证明交错鉴定法具有这种性质，即当 $n\to\infty$ 时，对 $s=q_1+1,\cdots,q$，有 $\hat{h}_s\to\infty$，对 $s=r_1+1,\cdots,r$，有 $\hat{\lambda}_s\to 1$。于是，渐近地，交错鉴定法能够自动地移除无关变量。

交错鉴定法的目标函数和在 4.4 节定义的基本一致，除了现在 $Y_i=\bar{g}(\bar{X}_i)+u_i$，$E(u_i|X_i)=0$，也就是说，条件均值函数只取决于相关的回归元 $\bar{x}$。m_i、$\hat{m}_i$ 和 $\hat{f}_i$ 的定义都没有变化，除非人们在 $g_i=g(X_i)$ 出现时需要用 $\bar{g}_i=\bar{g}(\bar{X}_i)$ 代替。

142

在 4.4 节中定义的 χ 函数修正的类似结果为：

$$\begin{aligned}
&\bar{\chi}(a_1,\cdots,a_{q_1},b_1,\cdots,b_{r_1})\\
&=\sum_{\bar{x}^d}\int\Big\{\Big(\sum_{s=1}^{r_1} b_s\sum_{\bar{v}^d} I_s(\bar{v}^d,\bar{x}^d)\{\bar{g}(\bar{x}^c,\bar{v}^d)-\bar{g}(\bar{x})\}\bar{f}(\bar{x}^c,\bar{v}^d)\bar{f}(\bar{x})^{-1}\\
&\quad+\frac{1}{2}\kappa_2\sum_{s=1}^{q_1} a_s^2\{\bar{g}_{ss}(\bar{x})+2\bar{f}_s(\bar{x})\bar{g}_s(\bar{x})\bar{f}(\bar{x})^{-1}\}\Big)^2\bar{f}(\bar{x})\\
&\quad+\sum_{\bar{x}^d}\int\frac{\kappa^q\bar{\sigma}^2(\bar{x})}{a_1\cdots a_{q_1}}\Big\}\bar{w}(\bar{x})\,\mathrm{d}\bar{x}^c
\end{aligned} \tag{4.37}$$

其中

$$\begin{aligned}
\bar{w}(\bar{x}^c,\bar{x}^d)&=\sum_{x_{r_1+1}^d,\cdots,x_r^d}\int\tilde{f}(x_{q_1+1}^c,\cdots,x_q^c,x_{r_1+1}^d,\cdots,x_r^d)\\
&\quad\times w(\bar{x}^c,x_{q_1+1}^c,\cdots,x_q^c)\,\mathrm{d}x_{q_1+1}^c,\cdots,\mathrm{d}x_q^c
\end{aligned}$$

横线（波浪线）符号是指只包含相关（不相关）回归元 $\bar{x}(\tilde{x})$ 的向量函数。$\bar{f}(\tilde{f})$ 是 $\bar{x}=(\bar{x}^c,\bar{x}^d)$ $(\tilde{x}=(\tilde{x}^c,\tilde{x}^d))$ 的联合概率密度，$\bar{m}_{ss}(\bar{f}_{ss})$ 是 $\bar{m}(\bar{f})$ 关于 $\bar{x}_s^c$ 的二阶导数，$\bar{m}=\bar{g}\bar{f}$。

与以前一样，用 $a_1^0,\cdots,a_{q_1}^0,b_1^0,\cdots,b_{r_1}^0$ 表示在它们都是非负的约束下最小化 $\bar{\chi}$ 的值。我们要求 a_s^0 和 b_s^0 是唯一被定义的，并且每一个都有界。

注意无关元素没有出现在 $\bar{\chi}$ 的定义中，这是因为无关的元素会在比值

$E[\hat{m}(x)]/E[\hat{f}(x)]$中相互抵消，我们把它记为 $\bar{\mu}_g(\bar{x})=E[\hat{m}(x)]/E[\hat{f}(x)]$。对相关变量 $\bar{x}$，我们要求：

$$\sum_{\bar{x}^d}\int_{\mathrm{supp}\ w}[\bar{\mu}_g(\bar{x})-\bar{g}(\bar{x})]^2\bar{w}(\bar{x})\bar{f}(\bar{x})\mathrm{d}\bar{x}^c \tag{4.38}$$

解释为 $h_1,\cdots,h_{q1},\lambda_1,\cdots,\lambda_{r1}$的一个函数，当且仅当这些平滑参数全部变为 0 时它也为 0。

由于我们应该让无关变量的平滑参数从 0 发散，我们将不再需要施加对所有的 s，有 $h_s=o(1)$，$\lambda_s=o(1)$的条件，而是给窗宽和核施加下面传统的条件。 143 定义：

$$H=\left(\prod_{s=1}^{q_1}h_s\right)\prod_{s=q_1+1}^{q}\min(h_s,1)$$

令 $0<\varepsilon<1/(q+4)$。假定

$$n^{\varepsilon-1}\leqslant H\leqslant n^{-\varepsilon}$$

$$\min(h_1,\cdots,h_q)>n^{-C},\max(h_1,\cdots,h_q)<n^{C},\text{某些 } C>0$$

核函数 K 是一个对称、紧支撑、Hölder 连续的概率密度函数

$$w(0)>w(\delta),\text{所有 }\delta>0 \tag{4.39}$$

如果 ε 任意小，上面关于 $h_1,\cdots,h_q$ 的条件基本上为当 $n\to\infty$ 时，$nh_1\cdots h_{q_1}\to\infty$，$h_1\cdots h_{q_1}\to 0$ 且 $h_1\cdots h_q\to 0$。另外，我们要求 h_s 趋于 0 的最快速度不能超过 n^{-C}，趋于无穷的最快速度不能超过 n^C（对无关变量）。

在这些条件下，我们得到下面的结果。

定理 4.8　在条件(4.28)、(4.34)、(4.38)和(4.39)下，令 $\hat{h}_1,\cdots,\hat{h}_q$，$\hat{\lambda}_1,\cdots,\hat{\lambda}_r$表示最小化 CV_r 的平滑参数，那么：

$$n^{1/(q_1+4)}\hat{h}_s\to a_s^0\text{依概率收敛},\ 1\leqslant s\leqslant q_1$$

$$P(\hat{h}_s>C)\to 1,\ q_1+1\leqslant s\leqslant q,\text{所有 }C>0$$

$$n^{2/(q_1+4)}\hat{\lambda}_s\to b_s^0\text{ 依概率收敛},\ 1\leqslant s\leqslant r_1$$

$$\hat{\lambda}_s\to 1\text{ 依概率收敛},\ r_1+1\leqslant s\leqslant r$$

$$n^{4/(q_1+4)}\inf\left[\mathrm{CV}_r(\hat{h},\hat{\lambda})-\frac{1}{n}\sum_i u_i^2M_i\right]\xrightarrow{p}\inf\bar{\chi} \tag{4.40}$$

定理 4.8 的证明在 Hall 等(2006)中给出。定理 4.8 说明了无关元素的平滑参数依概率收敛于各自的极下值；因此，所有的无关变量可以被(渐近地)自动平滑掉，而相关变量的平滑参数拥有无关变量没有出现时相同的最优性质。这与我们在 4.3 节讨论的只有离散回归元的情形下的结果形成鲜明对比，在那 144 里也存在无关回归元不会被平滑掉的可能性。

下一个定理给出了当用交错鉴定法选择平滑参数时，$\hat{g}(x)$的渐近正态

结果。

定理 4.9 在与定理 4.8 相同的条件下，令$\mathcal{S}^c=\text{supp }w, x=(x^c,x^d)\in\mathcal{S}^c\times\mathcal{S}^d$，那么：

$$(n\hat{h}_1\cdots\hat{h}_{q_1})^{1/2}\left[\hat{g}(x)-\bar{g}(\bar{x})-\sum_{s=1}^{q_1}\bar{B}_{1s}(\bar{x})\hat{h}_s^2-\sum_{s=1}^{r_1}\bar{B}_{2s}(\bar{x})\hat{\lambda}_s\right]$$

$$\xrightarrow{d} N\left(0,\frac{\kappa^{q_1}\bar{\sigma}^2(\bar{x})}{\bar{f}(\bar{x})}\right) \tag{4.41}$$

其中

$$\bar{B}_{1s}(\bar{x})=\frac{1}{2}\kappa_2\sum_{s=1}^{q_1}\left\{g_{ss}(\bar{x})+2\frac{\bar{f}_s(\bar{x})\bar{g}_s(\bar{x})}{\bar{f}(\bar{x})}\right\}$$

$$\bar{B}_{2s}(\bar{x})=\sum_{\bar{v}^d}\mathbf{1}_s(\bar{v}^d,\bar{x}^d)\{g(\bar{x}^c,\bar{v}^d)-\bar{g}(\bar{x})\}\bar{f}(\bar{x}^c,\bar{v}^d)\bar{f}(\bar{x})^{-1} \tag{4.42}$$

且$\bar{\sigma}^2(\bar{x})=E(u_i^2\mid\bar{X}_i=\bar{x})$。

定理 4.9 的证明在 Hall 等(2006)中给出。直观上我们可以这样理解，首先，注意当对$s=q_1+1,\cdots,q,\hat{h}_s\to\infty$和对$s=r_1+1,\cdots,r,\hat{\lambda}_s\to 1$时，我们有$\hat{g}(x,\hat{h}_1,\cdots,\hat{h}_q,\hat{\lambda}_1,\cdots,\hat{\lambda}_r)=\hat{g}(\bar{x},\hat{h}_1,\cdots,\hat{h}_{q_1},\hat{\lambda}_1,\cdots,\hat{\lambda}_{r_1})+(\text{s.o.})$。其次，由$\hat{h}_s-h_s^0=o_p(h_s^0)$，$(s=1,\cdots,q_1)$和$\hat{\lambda}_s-\lambda_s^0=o_p(\lambda_s^0)(s=1,\cdots,r_1)$，我们有$\hat{g}(\bar{x},\hat{h}_1,\cdots,\hat{h}_{q_1},\hat{\lambda}_1,\cdots,\hat{\lambda}_{r_1})=\hat{g}(\bar{x},\hat{h}_1^0,\cdots,\hat{h}_{q_1}^0,\hat{\lambda}_1^0,\cdots,\hat{\lambda}_{r_1}^0)+(\text{s.o.})$，所以定理 4.9 得证。

4.5.1 有序离散变量

到现在为止，我们把我们的注意力限制在离散变量是无序的情况。然而，如果一些离散变量是有序的(也就是有序分类变量)，那么我们将使用一个能够反映这些变量有序这个事实的核函数。假定x_s^d可以取c_s个不同的有序值$\{0,1,\cdots,c_s-1\}$，Aitchison 和 Aitken(1976, p. 29)建议当$|x_s^d-v_s^d|=j(0\leqslant s\leqslant c_s)$成立

145 时，使用由$l(x_s^d,v_s^d,\lambda_s)=\begin{bmatrix}c_s\\j\end{bmatrix}\lambda_s^j(1-\lambda_s)^{c_s-j}$给出的核函数，其中$\begin{bmatrix}c_s\\j\end{bmatrix}=c_s!/[j!(c_s-j)!]$。可以观察到这些权重加起来为 1；然而，当$c_s\geqslant 3$时，这个权重函数是有缺陷的，因为没有$\lambda_s$的值可以使$l(x_s^d,v_s^d,\lambda_s)$等于一个常数函数。因此，尽管$x_s^d$被证明是一个无关回归量，如果人们使用这个核，它绝不会被平滑掉。因此，我们建议使用如下另外一个核函数：

$$l(x_s^d,v_s^d,\lambda_s)=\lambda_s^{|x_s^d-v_s^d|} \tag{4.43}$$

λ_s的取值范围是$[0,1]$。再一次，当λ_s取其极大值($\lambda_s=1$)时，我们看到对于x_s^d的所有取值，$l(x_s^d,v_s^d,\lambda_s)\equiv 1$成立，$v_s^d\in\{0,1,\cdots,c_s-1\}$。在这种情形下，

x_s^d 在最后的估计中完全被平滑掉。

可以容易地证明当一些离散变量有序时，如果人们使用式(4.43)所给定的核函数，那么，上面提到的所有结果继续成立。

当然，如果一个有序离散变量取非常多不同的值，人们可以简单地把它看成连续变量去处理。在实践中这样得出的估计结果与人们把这个变量当作一个有序离散变量处理并使用式(4.43)给出的核函数产生的估计结果是类似的。

Bierens(1983,1987)以及 Ahmad 和 Cerrito(1994)也考虑了对一个回归函数的非参数估计把离散和连续变量都进行平滑。但是他们没有研究数据驱动的平滑参数选择这一基本问题。正如我们在这一节所展示的，用自动的数据驱动方法来选择平滑参数是重要的。此外，最小二乘交错鉴定法有它可以自动地(渐近地)消除不相关的解释变量这个特别有吸引力的性质。在有离散和连续变量的非参数背景中，最小二乘交错鉴定法单独成立，其他方法没有这个性质。

4.6 应　用

4.6.1 非家用食品消费

在 20 世纪 80 年代之前，在中国食品部门所增加的价值中非家用食品(FAFH)所占的比重是微不足道的。大部分膳食的家庭生产是使用家庭生产的谷物、生蔬菜和肉类，在国营的食品店购买或者直接从农民手中购买。20 世纪 80 年代之后，中国消费者在饭馆、餐厅、食堂的用餐日益增加，反映了中国收入快速增长。FAFH 在总食品支出中所占的份额从 1992 年的 5.03% 稳步增长到 2000 年的 14.70%。在 2000 年，人均每年 FAFH 支出达到 288 元，在中国城市总 FAFH 支出达到 1 320 亿元(159 亿美元)。在下面的应用中，可以看到非参数方法预期中国 FAFH 支出会继续增长，原因是中产阶级的增长，快速的城市化以及与美国(2001 年 40.3%，美国经济研究局，USDA)、加拿大(2001 年 35.6%，加拿大统计局)和其他发达国家(见 Jensen 和 Yen(1996))相比较低的 FAFH 份额。这个结论与用于为这种支出建模的一个流行的线性模型所得到的结论相反。

146

数据来源于中国国家统计局分别在 1992 年和 1998 年对城市居民的调查。因变量是人均家庭的 FAFH 支出(1992 年人民币)。解释变量包括家庭人均收入、家庭的大小、户主的受教育水平(七个类别)[①]、户主的年龄、一个 0-1 虚拟变量表明家庭是否在大城市和一个户主性别的虚拟变量。在这个研究中总样本

① 1. 小学以下，2. 小学，3. 初中，4. 高中，5. 中专，6. 两年大专，7. 学士及以上。

的大小1992年是3 459,1998年是3 359。Min, Fang和Li(2004)也估计了一个流行的线性模型用于比较。他们给出了一些图说明FAFH支出与收入之间的关系(所有其他回归元固定在样本中位数不变)。1992年和1998年的非参数FAFH支出曲线都表明在已经达到一个高收入水平后FAFH消费停止增长。这与参数模型的结果非常不同,在那里线性模型预测家庭收入越高FAFH消费越高。为了去核查哪种结果提供了对数据更好的描述,Min等分别计算了参数和非参数模型的拟合优度(R^2)。1992年和1998年数据线性模型的R^2值分别为0.128和0.170,而非参数模型中分别为0.382和0.348。

147

表4.1给出了不同收入水平的平均FAFH支出,以及用参数(普通最小二乘,OLS)和非参数模型预测的均值。高收入的家庭确实在FAFH上花费比那些收入略低的家庭要少。在1992年,收入超过5 000元的家庭比那些收入在3 000—5 000元的家庭在FAFH上花费少了11元。在1998年,收入超过9 000元的家庭比收入在4 000—9 000元的家庭在FAFH上花费少了52元。我们观察到对中低收入水平,参数和非参数模型都很好地预测了FAFH的平均支出。然而,对于高收入水平,参数模型给出了误导性的预测,而非参数方法在这种情形中做得更好。

表4.1 依收入分类的FAFH平均支出

1992年			
收入	数据平均	OLS平均	非参数平均
3 000元以下	82.78	82.12	83.26
3 000—5 000元	154.9	152.0	148.0
5 000元以上	143.9	221.7	144.9
1998年			
收入	数据平均	OLS平均	非参数平均
4 000元以下	119.40	120.1	121.9
4 000—9 000元	257.0	250.0	254.0
9 000元以上	205.2	441.1	232.9

表4.2报告了收入弹性的均值和中位数的估计值。在非参数模型中,家庭i的弹性η_i由$\eta_i = \frac{\partial \hat{g}(x_i, z_i)}{\partial x_{1i}} \frac{x_{1i}}{y_i}$计算得出,其中$x_{1i}$表示第$i$个家庭的收入,然而在线性模型中,$\eta_i = \beta_1 \frac{x_{1i}}{y_i}$。表4.2揭示了一些有趣的现象。比较1992年与1998年的弹性,非参数模型的结果表明大城市和中小城市家庭收入弹性的均值和中位数都增加了。相反,参数模型结果却表明只有中小城市家庭弹性的均值

增加了，而大城市家庭弹性的均值以及大城市和中小城市家庭弹性的中位数从 1992 到 1998 年都是下降的。这种互相冲突的估计结果可能是由于线性模型的错误设定导致的。正如我们之前讨论的，线性模型过高估计了高收入家庭的 FAFH 支出，因为它在 FAFH 消费中强加了一个不正确的线性收入(趋势)成分。错误设定的线性模型也过高估计了高收入家庭的弹性，从而导致了从 1992 年到 1998 年弹性的中位数下降的错误预测。

表 4.2　收入弹性的均值和中位数(η_i 代表家庭 i 的收入弹性)

参数结果	大城市		中小城市	
	1992 年	1998 年	1992 年	1998 年
均值	0.878	0.848	1.274	1.351
中位数	0.848	0.832	1.074	0.994
$\eta_i>1$ 的百分比	18.5%	13.9%	69.6%	48.8%
非参数结果	**大城市**		**中小城市**	
	1992 年	1998 年	1992 年	1998 年
均值	0.626	0.826	0.900	0.947
中位数	0.751	0.851	0.938	0.960
$\eta_i>1$ 的百分比	35.0%	39.9%	45.9%	45.8%

4.6.2　罢工量建模

我们考虑一个用于为 18 个经济合作与发展组织(OECD)国家罢工水平建模的年度观测值的面板数据。数据包括罢工量水平的观测值(每 1 000 个工薪收入者由于劳资纠纷所造成的损失的天数)和从 1951 年到 1985 年在 18 个 OECD 国家中的解释变量。不同国家间的罢工量的平均水平和方差差异非常大。数据分布也以一个长右尾和几个罢工量的大值为特征。我们使用下面的解释变量：(1) 国家代码；(2) 年份；(3) 罢工量；(4) 失业率；(5) 通货膨胀率；(6) 社会民主和劳工组织的议会代表；(7) 工会集中度的一个时间不变指标。数据是公开可得的(见 StatLib, http://lib.stat.cmu.edu)。因为一个国家的数据不完全，我们仅对 17 个有完整数据的国家进行分析。　148

Western(1996)分析了这些数据，他考虑了一个具有国家特定固定效应和时间趋势的线性面板数据模型。我们考虑一个非参数模型，它把国家代码处理为分类数据而其他的回归元是连续的。为了评估每个模型的表现，我们估计了每个模型，数据是 1951—1983 年间，然后，基于样本外预测的表现用 1984—　149

1985 年间的数据评估每个模型,并用预测平方误差作为我们的标准。我们以考虑交错鉴定窗宽开始,结果见表 4.3。

表 4.3　使用训练数据不同方法下的标准差和窗宽,$n_1=561$

	1	2	3	4	5	6
$\hat{\sigma}$	9.53	2.84	4.75	13.31	0.31	—
$\hat{h}/\hat{\lambda}$	102 565 846.11	4 800 821.61	5.84	30.56	408 328.03	0.12

注:(1) 年份;(2) 失业率;(3) 通货膨胀率;(4) 议会代表;(5) 工会集中度;(6) 国家代码。

正如从表 4.3 中所看见的,连续回归元年份、失业率以及工会集中度在相应的非参数估计中被有效平滑掉了。这意味着连续回归元通货膨胀、议会代表和离散回归元国家代码在上文所述的意义上是相关的。

我们接下来比较每个模型的样本外预测表现。参数面板模型(所有变量线性地进入模型)的相对均方误差相对于非参数交错鉴定法是 1.33。我们注意到不同的预测期对模型的相对预测表现几乎没有影响。我们也尝试了一个有交互项的参数模型(连续回归元的平方),相应的样本外预测均方误差甚至比线性模型的均方误差预测大(它的预测均方误差与非参数交错鉴定法的比值是 1.44),然而一个只有一个常数、失业率和通货膨胀率的简单参数模型,它的预测均方误差比非参数交错鉴定法大 15%(相对预测均方误差是 1.15),这意味着一个线性参数的设定是不合适的。

150

4.7　习　题

习题 4.1　推导下式给出的非平滑估计量 $p(x)$ 的均方误差

$$\tilde{p}(x)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i=x)$$

其中 $\mathbf{1}(\cdot)$ 是一个定义如下的示性函数:

$$\mathbf{1}(X_i=x)=\begin{cases}1, & X_i=x\\ 0, & \text{其他}\end{cases}$$

其中 $X_i\in\mathcal{S}=\{0,1,\cdots,c-1\}$ 是一个有有限支撑的离散随机变量。

习题 4.2　考虑由下式给出的核估计量 $p(x)$:

$$\hat{p}(x)=\frac{1}{n}\sum_{i=1}^{n}L(X_i,x,\lambda)$$

其中 x 是标量,$L(\cdot)$ 是一个定义如下的核函数:

$$L(X_i,x,\lambda) = \begin{cases} 1-\lambda, & X_i = x \\ \lambda/(c-1), & 其他 \end{cases}$$

其中 $\lambda \in [0,(c-1)/c]$。

(i) 推导估计量的偏误，然后证明如果 X 有一个（离散的）均匀分布，则对所有的 X 有 $p(x)=1/c$，那么核估计量 $\hat{p}(x)$ 对任意可接受的 λ 值是无偏的。

(ii) 证明

$$\mathrm{var}(\hat{p}(x)) = \frac{p(x)(1-p(x))}{n}\left(1-\lambda\frac{c}{c-1}\right)^2$$

证明如果 X 有一个（离散的）均匀分布，则对所有的 X，$p(x)=1/c$，当 $\lambda=(c-1)/c$ 即它的上界时，$\hat{p}(x)$ 的方差为 0。

(iii) 给定上述结论，当潜在分布是均匀分布时，核估计量 $\hat{p}(x)$ 是否可能有一个为零的均方误差？当 λ 满足什么条件时这种情况会发生？证明你的结论。

习题 4.3

151

(i) 证明 $E[\hat{p}(x^d)]$ 和 $\mathrm{var}(\hat{p}(x^d))$ 确实由式(4.5)给出，其中

$$B_{p,s} = \sum_{z^d}\frac{1}{c_s-1}\mathbf{1}_s(x^d,z^d)p(z^d) - p(x^d)$$

其中 $\mathbf{1}_s(x^d,z^d)$ 的定义见式(4.24)。

(ii) 证明 $E(\hat{m}(x^d)^2) = O\left(\sum_{s=1}^r \lambda_s^2 + n^{-1}\right)$，其中 $\hat{m}(x^d)$ 在式(4.15)中定义。注意 $O\left(\left(\sum_{s=1}^r \lambda_s\right)^2\right) = O\left(\sum_{s=1}^q \lambda_s^2\right)$。

习题 4.4　Ω_r 是一个 $r\times r$ 阶矩阵，其中第 (s,t) 个元素为

$$\sum_{x^d\in\mathcal{S}^d}[p(x^d)-p_{1,s}(x^d)][p(x^d)-p_{1,t}(x^d)]$$

其中

$$p_{1,s}(x^d) = \frac{1}{c_s-1}\sum_{z^d\in\mathcal{S}^d}p(z^d)\mathbf{1}_s(x^d,z^d) \tag{4.44}$$

注意 $p_{1,s}(x^d)$ 也是一个概率指标，因为 $p_{1,s}(x^d)$ 是在 c_s-1，$z^d\in\mathcal{S}^d$ 上的平均概率，与 x^d 的不同仅在于第 s 个元素的不同，即 $p_{1,s}(x^d) = \frac{1}{c_s-1}\sum_{z_s^d\neq x_s^d}p(z_s^d,x_{-s}^d)$。

考虑 $r=2$ 的简单情形。那么：

$$\lambda'_{(2)}\Omega_2\lambda_{(2)} = \sum_{x^d}\{\lambda_1[p(x^d)-p_{1,1}(x^d)] + \lambda_2[p(x^d)-p_{1,2}(x^d)]\}^2$$

证明当且仅当存在 $x^d, z^d\in\mathcal{S}^d$ 使得 $p(x^d)\neq p_{1,1}(x^d)$ 且 $p(z^d)\neq p_{1,2}(z^d)$ 时，Ω_2 是正定的。

习题 4.5 使用式(4.3),我们有:

$$I_{1n} = \sum_{x^d} [\hat{p}(x^d)]^2 = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \sum_{x^d} L_{\lambda,ix} L_{\lambda,jx} = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} L_{\lambda,ij}^{(2)}$$

其中 $L_{\lambda,ij}^{(2)} = \sum_{x^d} L_{\lambda,ix} L_{\lambda,jx}$。

因此,我们有

$$\mathrm{CV}_p(\lambda) = n^{-2} \sum_{i=1}^{n} L_{\lambda,ii}^{(2)} + \frac{1}{n(n-1)} \sum_{i} \sum_{j \neq i} [L_{\lambda,ij}^{(2)} - 2L_{\lambda,ij}]$$
$$- \frac{1}{n^2(n-1)} \sum_{i} \sum_{j \neq i} L_{\lambda,ij}^{(2)} \tag{4.45}$$

令 $\mathrm{CV}_p(\lambda)$ 在式(4.45)中定义,并假定 x^d 是一元的且从$\{0,1,\cdots,c-1\}$中取值,证明(λ 是一个标量)

$$E(\mathrm{CV}_p(\lambda)) = D_1\lambda^2 - D_2\lambda n^{-1} + o(\lambda^2 + \lambda n^{-1})$$
$$+ \text{独立于 } \lambda \text{ 的项} \tag{4.46}$$

其中 D_j 为(正的)常数。推导 D_j 的显性表达式。

注意关于 λ 最小化式(4.46),得出 $\tilde{\lambda} = n^{-1}D_2/(2D_1) = O(n^{-1})$。

习题 4.6

(i) 证明定理 4.2(i)。

提示:利用定理 4.1 和式(3.3)的结论。

(ii) 证明定理 4.2(ii)。

提示:利用定理 4.1 和式(3.4)的结论。

习题 4.7 在 x^d 为一元的情形下证明定理 4.4。

提示:当 x^d 为一元时,有 $L_{ij,\lambda} = \mathbf{1}_{d_{ij}=0} + \lambda\mathbf{1}_{d_{ij}\neq 0}$($d_{ij} = |x_i^d - x_j^d|$)。记:

$$\hat{p}_{-i}(x_i^d) = n^{-1} \sum_{j \neq i} [\mathbf{1}_{d_{ij}=0} + \lambda\mathbf{1}_{d_{ij}\neq 0}] \equiv \hat{p}_{i0} + \lambda\hat{p}_{i1}$$

那么

$$\frac{1}{\hat{p}_{-i}(x_i^d)} = \frac{1}{\hat{p}_{i0}}\left[1 - \lambda\frac{\hat{p}_{i1}}{\hat{p}_{i0}}\right] + O(\lambda^2)$$

可以把 $\mathrm{CV}_d(\lambda)$ 扩展为一个关于 λ 的幂级数,即:

$$\mathrm{CV}_d = A_1\lambda^2 + A_{2n}\lambda n^{-1} + o(\lambda n^{-1} + \lambda^2) + \text{与 } \lambda \text{ 无关的项}$$

其中 $A_1 > 0$ 是一个正的常数,A_{2n} 是一个 $O_p(1)$ 的随机变量。

习题 4.8 在证明定理 4.5 时,我们用到 $\hat{\sigma}^2(x^d) = \sigma(x^d) + o_p(1)$,证明这一结论。

提示:记 $\hat{\sigma}^2(x^d) - \sigma^2(x^d) = [\hat{\sigma}^2(x^d) - \sigma^2(x^d)]\hat{p}(x^d)/\hat{p}(x^d)$,使用 $\hat{\lambda}_s = O_p(n^{-1})$。根据大数定律证明$\hat{\sigma}^2(x^d)\hat{p}(x^d) = n^{-1}\sum_{i=1}^{n} u_i^2 \mathbf{1}(X_i^d = x^d) + O_p(n^{-1}) \xrightarrow{p}$

$\sigma^2(x^d)p(x^d)$。

习题 4.9 从式(4.23)推出式(4.21)。

习题 4.10 a_s^0 和 b_s^0 被定义为满足最小化定义在式(4.31)中的χ的唯一有界的值,考虑 $q=1$ 和 $r=1$ 的情形,通过最小化式(4.31)解出 a_1^0 和 b_1^0。 153

提示:注意$\chi(a_1,b_1)=C_1a_1^4+C_2b_1^2-C_3a_1^2b_1+C_0/a_1$,其中对 $j=0,1,2$,$C_j>0$。这可以写成:

$$\chi(a_1,b_1)=C_2[b_1-A_1a_1^2]^2+A_2a_1^4+C_0/a_1$$

其中 $A_1=C_3/(2C_2)$,$A_2=C_1-C_3^2/(4C_2)>0$。因此,我们有 $b_1^0=\max\{0,A_1(a_1^0)^2\}$(因为 b_1^0 非负),如果 $A_1\geqslant 0$,有 $a_1^0=[C_0/(4A_2)]^{1/5}$,如果 $A_1<0$,有 $a_1^0=[C_0/(4C_1)]^{1/5}$。

习题 4.11 考虑 $r=1$ 的简单情形,证明当且仅当 $g(x^d)$不是一个常数函数(这里 x^d 是一元的)时,假定 4.1 成立。

提示:对所有的 $x^d\in S^d$(x^d 为一元的,因为 $r=1$),假定 4.1 可变化为 $\sum_{z^d\in S^d}\{[g(x^d)-g(z^d)][\mathbf{1}(x^d=z^d)+\lambda\mathbf{1}(x^d\neq z^d)]\}^2=0$,这等价于 $\lambda^2\sum_{z^d\in S^d}[g(x^d)-g(z^d)]^2\mathbf{1}(x^d\neq z^d)=0$,因为$[g(x^d)-g(z^d)]\mathbf{1}(x^d=z^d)\equiv 0$。然而,既然 $g(x^d)$ 不是一个常数函数,我们知道 $\sum_{z^d\in S^d}[g(x^d)-g(z^d)]^2\mathbf{1}(x^d\neq z^d)>0$。因此,我们必须有 $\lambda=0$。于是,当 x^d 是一元时,当且仅当 $g(x^d)$ 不是一个常数函数时,假定 4.1 成立。

第5章　条件密度估计

155

条件概率密度函数构成了今天使用的大部分流行统计方法的支柱,尽管通常它们不在一个参数框架下进行直接建模。在核框架下它们甚至更少受到关注。然而,正如我们即将看到的,这种方法在一系列任务中是极其有用的,包括计数数据建模或者预测消费者的选择(见 Cameron 和 Trivedi(1998)对计数数据模型的一个全面的处理)。本章中我们将讨论条件概率密度函数的非参数估计。我们着重关注实际上相关的离散和连续数据混合的情形。我们在第 3 章和第 4 章已经提到,用一种特别的方式平滑离散变量相对于使用基于频率的方法具有理论和实践上的优势,因此下文我们将直接进行到平滑离散和连续变量。注意,我们可能把条件变量作为回归元,尽管我们是在估计条件密度而不是回归函数。

5.1　条件密度估计:相关变量的情形

令 $f(\cdot)$ 和 $\mu(\cdot)$ 分别表示 (X,Y) 的联合密度以及 X 的边际密度。在下面我们将把 Y 作为一个因变量(即 Y 是被解释变量),X 是协变量(即 X 是解释变量)。我们

156

用 $\hat{f}$ 和 $\hat{\mu}$ 表示其核估计。我们通过下式估计条件密度 $g(y|x)=f(x,y)/\mu(x)$:

$$\hat{g}(y\mid x)=\hat{f}(x,y)/\hat{\mu}(x) \tag{5.1}$$

我们首先考虑 Y 是一元连续随机变量的情形,然后讨论 Y 是一元离散变量的情形。我们在 5.4 节处理 Y 是多元的情况。我们对 $f(\cdot)$ 和 $\mu(\cdot)$ 的估计是:

$$\hat{f}(x,y) = n^{-1}\sum_{i=1}^{n} K_\gamma(x,X_i)k_{h_0}(y,Y_i) \tag{5.2}$$

$$\hat{\mu}(x) = n^{-1}\sum_{i=1}^{n} K_\gamma(x,X_i) \tag{5.3}$$

其中 h_0 是 Y 的平滑参数，$\gamma=(h,\lambda)$，且

$$K_\gamma(x,X_i) = W_h(x^c,X_i^c)L(x^d,X_i^d,\lambda)$$

$$W_h(x^c,X_i^c) = \prod_{s=1}^{q}\frac{1}{h_s}w\left(\frac{x_s^c - X_{is}^c}{h_s}\right)$$

$$L(x^d,X_i^d,\lambda) = \prod_{s=1}^{r}[\lambda_s/(c_s-1)]^{N_{is}(x)}(1-\lambda_s)^{1-N_{is}(x)}$$

$N_{is}(x)=\mathbf{1}(X_{is}^d\neq x_s^d)$ 是一个示性函数，当 $X_{is}^d\neq x_s^d$ 时它等于1，否则等于0，$k_{h0}(y,Y_i)=h_0^{-1}k((y-Y_i)/h_0)$。

正如我们在前面章节所强调的，在应用背景中需要数据驱动的窗宽选择方法，如交错鉴定法。下面我们讨论两种不同的交错鉴定法。第一种是基于最小化 $\hat{g}(y|x)$ 和 $g(y|x)$ 之差的加权积分平方，第二种方法是似然方法。正如我们将要看到的，第一种方法有一些理想的最优性质，但是在大样本中是计算繁重的。第二种方法计算代价更小，但是在连续变量的分布有厚尾时会导致不一致的估计。

5.2 条件密度窗宽选择

157

5.2.1 最小二乘交错鉴定：相关变量的情形

与非条件密度估计的情形一样，当用核方法估计条件密度时，我们可以使用最小二乘交错鉴定法来选择平滑参数。我们考虑下面基于一个加权积分平方误差（$\int dx = \sum_{x^d}\int dx^c$）的准则，

$$\begin{aligned}\text{ISE} &= \int\{\hat{g}(y\mid x)-g(y\mid x)\}^2\mu(x)M(x^c)\,dx\,dy \\ &= I_{1n}-2I_{2n}+I_{3n}\end{aligned} \tag{5.4}$$

其中 $M(\cdot)$ 是一个权重函数，

$$I_{1n} = \int\hat{g}(y\mid x)^2\mu(x)M(x^c)\,dx\,dy,\quad I_{2n} = \int\hat{g}(y\mid x)f(x,y)M(x^c)\,dx\,dy$$

另外，$I_{3n}=\int g^2(y\mid x)\mu(x)M(x^c)\,dx\,dy$ 不依赖于用于计算 $\hat{f}$ 和 $\hat{\mu}$ 的平滑参数。我们观察到

$$I_{1n} = \int \hat{G}(x) \frac{\mu(x)}{\hat{\mu}(x)^2} M(x^c)\,\mathrm{d}x = E_X\Big[\frac{\hat{G}(x)}{\hat{\mu}(x)^2} M(x^c)\Big]$$

这里的期望是关于 X 的,而不是随机观测值 $\{X_i, Y_i\}_{i=1}^n$,它们被用于计算 $\hat{G}(\cdot)$ 和 $\mu(\cdot)$,$\hat{G}(x) = \int \hat{f}(x,y)^2\mathrm{d}y$,这里的 $\hat{G}(\cdot)$ 被定义为:

$$\hat{G}(x) = \frac{1}{n^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n} K_\gamma(x, X_{i_1}) K_\gamma(x, X_{i_2}) \int w_{h_0}(y, Y_{i_1}) w_{h_0}(y, Y_{i_2})\,\mathrm{d}y$$

类似地,I_{2n}可以被写成:

$$E_Z[\hat{g}(Y \mid X) M(X^c)] = E_Z[\hat{f}(Y,X) M(X^c)/\hat{\mu}(X)]$$

这里的期望是关于 $Z=(Y,X)$ 的。因此,下面采用交错鉴定法得到对 I_{1n}和 I_{2n}的近似值 $\hat{I}_{1n}$和 $\hat{I}_{2n}$:

$$\hat{I}_{1n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{G}_{-i}(X_i) M(X_i^c)}{\hat{\mu}_{-i}(X_i)^2}, \quad \hat{I}_{2n} = \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{f}_{-i}(X_i, Y_i) M(X_i^c)}{\hat{\mu}_{-i}(X_i)}$$

158 其中下标 $-i$ 表示去一估计量,例如:

$$\hat{G}_{-i}(X_i) = \frac{1}{(n-1)^2}\sum_{i_1=1, i_1\neq i}^{n}\sum_{i_2=1, i_2\neq i}^{n} K_n(X_i, X_{i_1}) K_n(X_i, X_{i_2}) \times \int w_{h_0}(y, Y_{i_1}) w_{h_0}(y, Y_{i_2})\,\mathrm{d}y$$

因此,我们的交错鉴定目标函数为:

$$\mathrm{CV}_g(h_0, h, \lambda) = \hat{I}_{1n}(h_0, h, \lambda) - 2\hat{I}_{2n}(h_0, h, \lambda)$$

其中$(h, \lambda) = (h_1, \cdots, h_q, \lambda_1, \cdots, \lambda_r)$。

根据在 4.4.2 节的提出分析思路,再次使用 $\hat{\mu}(x)^{-1} = \mu(x)^{-1} + (\mu(x) - \hat{\mu}(x))/[\mu(x)\hat{\mu}(x)]$来处理随机分母的出现,容易看出 $\hat{g}(y|x) - g(y|x)$ 的首项是$[\hat{g}(y|x) - g(y|x)]\hat{\mu}(x)/\mu(x) = [\hat{f}(x,y) - \hat{\mu}(x) g(y|x)]/\mu(x)$,Hall 等(2004)进一步证明了 CV_g 的首项是:

$$\mathrm{CV}_{g0} = \int \{E[\hat{f}(x,y) - \hat{f}(x) g(y \mid x)]^2 M(x^c)/\mu(x)^2\}\,\mathrm{d}x\mathrm{d}y \qquad (5.5)$$

令 $\rho_s = \lambda_s/\{(1-\lambda_s)(c_s - 1)\}$,用 $f_{00}(x^c, x^d, y)$ 和 $f_{ss}(x^c, x^d, y)$ 表示 $f(x^c, x^d, y)$ 关于 y 和 x_s^c 的二阶偏导,我们可以写成:

$$\begin{aligned} E\{\hat{f}(x,y)\} &= \sum_{v^d}\Big\{\prod_{s=1}^{r}(1-\lambda_s)\rho_s^{1(v_s^d \neq x_s^d)}\Big\}\Big\{\int \prod_{s=1}^{q} k(z_s)\Big\} M(v) \\ &\quad \times f(x^c - hz, x^d, y - h_0 v)\,\mathrm{d}z_1 \cdots \mathrm{d}z_q \mathrm{d}v \\ &= f(x,y) + \sum_{s=1}^{r}\lambda_s\Big\{\frac{1}{c_s - 1}\sum_{v^d}\mathbf{1}_s(v^d, x^d) f(x^c, v^d, y) - f(x,y)\Big\} \end{aligned}$$

$$+\frac{1}{2}\kappa_2 h_0^2 f_{00}(x,y)+\frac{1}{2}\kappa_2\sum_{s=1}^{q}h_s^2 f_{ss}(x,y)+o(\eta_2) \tag{5.6}$$

其中 $\mathbf{1}_s(v^d,x^d)$ 在式(4.24)中定义，$\eta_2=\sum_{s=1}^{r}\lambda_s+\sum_{s=0}^{q}h_s^2$，且

159

$$E(\hat{\mu}(x))=\sum_{v^d}P(X^d=v^d)\{\prod_{s=1}^{r}(1-\lambda_s)\rho_s^{\mathbf{1}(v_s^d\neq x_s^d)}\}$$

$$\times\left\{\int\sum_{s=1}^{q}k(z_s)\right\}M(v)f(x^c-hz,x^d,y-h_0v)\,\mathrm{d}z_1\cdots\mathrm{d}z_q\mathrm{d}v$$

$$=\mu(x)+\sum_{s=1}^{r}\lambda_s\left\{\frac{1}{c_s-1}\sum_{v^d}\mathbf{1}_s(v^d,x^d)\mu(x^c,v^d)-\mu(x)\right\}$$

$$+\frac{1}{2}\kappa_2\sum_{s=1}^{q}h_s^2\mu_{ss}(x,y)+o(\eta_2) \tag{5.7}$$

因此，由式(5.6)和式(5.7)可以导出：

$$E\{\hat{f}(x,y)-\hat{\mu}(x)g(y\mid x)\}$$

$$=\prod_{s=1}^{r}(1-\lambda_s)\rho_s^{\mathbf{1}(v_s^d\neq x_s^d)}\left\{f(x^c,v^d,y)-\frac{\mu(x^c,v^d)}{\mu(x)}f(x,y)\right\}$$

$$+\frac{1}{2}\kappa_2 h_0^2 f_{00}(x,y)$$

$$+\frac{1}{2}\kappa_2\prod_{s=1}^{q}h_s^2\left\{f_{ss}(x,y)-\frac{\mu_{ss}(x)}{\mu(x)}f(x,y)\right\}+o(\eta_2) \tag{5.8}$$

注意：

$$\begin{aligned}\mathrm{var}(\hat{f}(x,y)-\hat{\mu}(x)g(y\mid x))&=n^{-1}\mathrm{var}[K_n(x,X_i)\{k_{h_0}(y,Y_i)-g(y\mid x)\}]\\&=n^{-1}E[\{K_n(x,X_i)k_{h_0}(y,Y_i)\}^2]+o(\eta_1)\\&=\kappa^{q+1}f(x,y)\eta_1+o(\eta_1)\end{aligned} \tag{5.9}$$

其中 $\eta_1=(nh_1\cdots h_q)^{-1}$。

联立式(5.8)和式(5.9)，我们可以得到

$$\mathrm{CV}_{g0}(h_0,h,\lambda)=n^{-q/(q+4)}\chi_g(a_0,a,b) \tag{5.10}$$

其中

160

$$\chi_g(a_0,a,b)=\sum_{x^d}\int\Bigg(\Bigg[\sum_{s=1}^{r}\frac{b_s}{c_s-1}\sum_{v^d}\mathbf{1}(v^d,x^d)$$

$$\times\left\{f(x^c,v^d,y)-\frac{\mu(x^c,v^d)}{\mu(x)}f(x,y)\right\}+\frac{1}{2}\kappa_2 a_0^2 f_{00}(x,y)$$

$$+\frac{1}{2}\kappa_2\sum_{s=1}^{q}a_s^2\left\{f_{ss}(x,y)-\frac{\mu_{ss}(x)}{\mu(x)}f(x,y)\right\}\Bigg]^2$$

$$+\frac{\kappa^{q+1}f(x,y)}{a_1\cdots a_q}\bigg)\frac{M(x^c)}{\mu(x)}\mathrm{d}x^c\mathrm{d}y \tag{5.11}$$

其中 a_s 和 b_s 分别通过 $h_s=n^{-1/(q+5)}a_s$ 和 $\lambda_s=b_s n^{-2/(q+5)}$ 来定义。

用 $a_0^0,a_1^0,\cdots,a_q^0,b_1^0,\cdots,b_r^0$ 表示最小化 χ_g 的那些值,我们假定 $a_0^0,a_1^0,\cdots,a_q^0,b_1^0,\cdots,b_r^0$ 是唯一确定的并且每个都是有界的。用 $h_1^0,\cdots,h_q^0,\lambda_1^0,\cdots,\lambda_r^0$ 表示最小化 CV_{g0} 的 $h_1,\cdots,\lambda_r$ 的值,那么 $h_s^0\sim n^{-1/(q+5)}a_s^0,\lambda_s^0\sim n^{-2/(q+5)}b_s^0$。同样,给定 $CV_g=CV_{g0}+(\text{s.o.})$,我们期望 $\hat{h}_s=h_s^0+(\text{s.o.}),\hat{\lambda}_s=\lambda_s^0+(\text{s.o.})$,我们在下一个定理中证明这确实成立。

定理 5.1

$$\begin{aligned}&n^{1/(q+5)}\hat{h}_s\xrightarrow{p}a_s^0,\quad s=0,1,\cdots,q\\&n^{2/(q+5)}\hat{\lambda}_s\xrightarrow{p}b_s^0,\quad s=1,\cdots,r\\&n^{4/(q+5)}\inf CV_g\to\inf\chi_g\text{依概率分布}\end{aligned}\tag{5.12}$$

式(5.12)的证明见 Hall 等(2004)。

运用定理 5.1,我们可以得到 $\hat{g}(y|x)$ 的渐近分布。我们将推迟讨论这个结论到 5.3 节,届时我们将讨论更一般的允许存在不相关协变量的情形。

5.5.2 最大似然交错鉴定:相关变量的情形

在 5.2.1 节讨论过的最小二乘交错鉴定方法有一些理想的最优性质,这些性质在定理 5.1 中已经给出。然而,最小二乘交错鉴定法在条件概率密度函数的估计中计算成本高,特别是当样本容量较大时。这是因为目标函数 CV_g(即 $\hat{I}_{1n}$)涉及三个求和。当样本容量较小时,可以通过把暂时权重矩阵存储在电脑内存中而实现显著的速度改善,而不是在计算 CV_g 时重新计算它们的各个部分。但是即使样本容量适中,存储这些权重矩阵的内存要求将很快超过大多数计算机的存储能力。这是一个典型的内存与速度计算上的权衡取舍。然而,相比之下,最大似然交错鉴定从计算角度而言有很大的优点,正如我们下面证明的。

161

似然交错鉴定方法选择 $h_1,\cdots,h_q,\lambda_1,\cdots,\lambda_r$ 来最大化对数似然函数

$$\mathcal{L}=\sum_{i=1}^{n}\ln\hat{g}_{-i}(Y_i\mid X_i) \tag{5.13}$$

其中 $\hat{g}_{-i}(Y_i|X_i)=\hat{f}_{-i}(X_i,Y_i)/\hat{m}_{-i}(X_i)$,$\hat{f}_{-i}(X_i,Y_i)$ 和 $\hat{m}_{-i}(X_i)$ 分别是 $f(X_i,Y_i)$ 和 $\mu(X_i)$ 的去一核估计量。目标函数(5.13)中所包含的求和比最小二乘交错鉴定少一个,因此计算负担较少。当连续变量来自厚尾分布时,似然交错鉴定

存在一个问题。在这种情况下,似然交错鉴定倾向于过度平滑数据,并且可能导致估计结果的不一致(见 Hall (1987a,1987b))。

两种实际的方法可以用于防止这个问题的发生。第一种,可以把似然交错鉴定的平滑参数与一些专门的公式(如 $h_s = x_{cd} n^{-1/(q+4)}$)相比较,如果这两种平滑参数是可以比较的,则不存在交错鉴定法的过度平滑问题。当最大似然交错鉴定过度平滑数据时,存在两种可能性:要么它可能会导致不一致的估计,要么相关变量是不相关的。我们注意到,在后一种情况中过度平滑实际上是很理想的。为了区分这两种情况,人们可以用第二种方法去比较基于似然交错鉴定的样本外预测与基于参数方法的预测。如果对所提供的数据,非参数方法表现更好或者与相应参数方法相当,这说明似然交错鉴定可能没有成为潜在的不一致问题的受害者。当然,如果最小二乘交错鉴定法计算是可行的,最小二乘交错鉴定平滑参数也可以被计算出来,那么它们也可以用来判断似然交错鉴定是否产生了不合理的偏大的平滑参数。 162

5.3 条件密度估计:无关变量的情形

我们现在考虑有一些协变量可能是不相关的情形。使用 4.5 节定义的符号,对整数 $0 \leq q_1, q_2 \leq q, 0 \leq r_1, r_2 \leq r$,满足 $q_1 + q_2 = q, r_1 + r_2 = r$,令 $\bar{X}$ 由 X^c 的前 q_1 个元素和 X^d 的前 r_1 个元素组成,$\tilde{X} = X/\bar{X}$ 表示 X 的剩余元素,我们假定:

$$(Y, \bar{X}) \text{ 独立于 } \tilde{X} \tag{5.14}$$

条件(5.14)意味着 $g(y|\bar{x}, \tilde{x}) = g(y|\bar{x})$,即无关变量 $\tilde{x}$ 不影响 $g(y|\bar{x})$。再一次,我们没有事先假定该信息已知,实际上我们估计的是 $g(y|x)$ 而不是 $g(y|\bar{x})$。我们将证明通过用交错鉴定法选择平滑参数,无关变量将被(渐近地)平滑掉。

函数 χ_g 现在需要被修改为:

$$\begin{aligned}
\bar{\chi}_g(a_0, a, b) = \sum_{\bar{x}^d} \int \Bigg(\Bigg[\sum_{s=1}^{r_1} \frac{b_s}{c_s - 1} \sum_{\bar{v}^d} \mathbf{1}_s(\bar{v}_s^d, \bar{x}_s^d) \\
\times \left\{ \bar{f}(\bar{x}^c, \bar{v}^d, y) - \frac{\bar{\mu}(\bar{x}^c, \bar{v}^d)}{\bar{\mu}(\bar{x})} \bar{f}(\bar{x}, y) \right\} + \frac{1}{2} \kappa_2 a_0^2 \bar{f}_{00}(\bar{x}, y) \\
+ \frac{1}{2} \kappa_2 \sum_{s=1}^{q_1} a_s^2 \left\{ \bar{f}_{ss}(\bar{x}, y) - \frac{\bar{\mu}_{ss}(\bar{x})}{\bar{\mu}(\bar{x})} \bar{f}(\bar{x}, y) \right\} \Bigg]^2 \\
+ \frac{\kappa^{q_1+1} \bar{f}(\bar{x}, y)}{a_1 \cdots a_{q_1}} \Bigg) \frac{\bar{w}(\bar{x})}{\bar{\mu}(\bar{x})} d\bar{x}^c dy
\end{aligned} \tag{5.15}$$

其中

$$\bar{w}(\bar{x}^c,\bar{x}^d) = \sum_{x^d_{r_1+1},\cdots,x^d_r}\int \frac{w(\bar{x}^c,x^c_{q_1+1},\cdots,x^c_q)}{f(\bar{x}^c,x^c_{q_1+1},\cdots,x^c_q,\bar{x}^d,x^d_{r_1+1},\cdots,x^d_r)}\mathrm{d}x^c_{q_1+1}\cdots\mathrm{d}x^c_q$$

163 我们从定理4.8中知道对于非参数回归,最小二乘交错鉴定渐近地平滑掉了无关变量。下一个定理证明了对交错鉴定条件密度估计而言,类似的结论依然成立。

定理5.2 在类似于定理4.8中给出的条件下,令$\hat{h}_0,\hat{h}_1,\cdots,\hat{h}_q,\hat{\lambda}_1,\cdots,\hat{\lambda}_r$表示最小化$\mathrm{CV}_g$的平滑参数,那么

$$n^{1/(q_1+5)}\hat{h}_s \xrightarrow{p} a^0_s, 0 \leqslant j \leqslant q_1$$

$$P(\hat{h}_s > C) \to 1, q_1+1 \leqslant s \leqslant q, \text{所有 } C > 0$$

$$n^{2/(q_1+5)}\hat{\lambda}_s \xrightarrow{p} b^0_s, 1 \leqslant s \leqslant r_1$$

$$\hat{\lambda}_s \xrightarrow{p} \frac{c_s-1}{c_s}, r_1+1 \leqslant j \leqslant r$$

$$n^{4/(q_1+5)}\inf \mathrm{CV}_r(\hat{h},\hat{\lambda}) \xrightarrow{p} \inf \bar{\chi} \tag{5.16}$$

Hall等(2004)给出了定理5.2的证明,类似于定理4.8,定理5.2说明了交错鉴定平滑参数选择可以渐近地移除不相关的条件变量。下一个定理给出了$\hat{g}(y|x)$的渐近正态分布。

定理5.3 用交错鉴定平滑参数$\hat{h}_0,\cdots,\hat{h}_q,\hat{\lambda}_1,\cdots,\hat{\lambda}_r$计算$\hat{g}(y|x)$,那么

$$(n\hat{h}_1\cdots\hat{h}_{q_1})^{1/2}\Big(\hat{g}(y\mid x) - g(y\mid x) - \sum_{s=0}^{q_1}B_{1s}(\bar{x},y)\hat{h}^2_s - \sum_{s=1}^{r_1}B_{2s}(\bar{x},y)\hat{\lambda}_s\Big)$$

$$\to N(0,\sigma^2_g(\bar{x},y)) \text{ 依分布收敛} \tag{5.17}$$

其中

$$B_{10}(\bar{x},y) = \frac{1}{2}\kappa\hat{h}^2_0\bar{g}_{00}(y\mid\bar{x})$$

$$B_{1s}(\bar{x},y) = \frac{1}{2}\kappa_2\left\{\frac{\bar{f}_{ss}(y\mid\bar{x})}{\bar{\mu}(\bar{x})} - \frac{\bar{\mu}_{ss}(\bar{x})}{\bar{\mu}(\bar{x})}\bar{g}(y\mid\bar{x})\right\}, s = 1,\cdots,q_1$$

$$B_{2s}(\bar{x},y) = \frac{1}{c_s-1}\sum_{\bar{v}^d}\mathbf{1}_s(\bar{v}^d,x^d_s)\left\{\bar{g}(y\mid\bar{x}^c,\bar{v}^d) - \frac{\bar{\mu}(\bar{x}^c,\bar{v}^d)}{\bar{\mu}(\bar{x})}\bar{g}(y\mid\bar{x})\right\}$$

$$\sigma^2_g(\bar{x},y) = \kappa^{q_1+1}\bar{g}(y\mid\bar{x})/\bar{\mu}(\bar{x})$$

164 到目前为止我们都是假定Y是一个连续随机变量。现在我们转向关注Y是离散的情形。如果Y取c_0个不同的值,那么我们可以用分类核$l(y,Y_i,\lambda_0) = \lambda_0^{N_i(y)}(1-\lambda_0)^{1-N_i(y)}(N_i(y) = 1(Y_i \neq y))$来代替连续核$k_{h0}(y,Y_i) = h_0^{-1}k((y-$

$Y_i)/h_0$)。定理5.2将被修改，用 q_1+4 代替 q_1+5，用 $n^{2/(q1+4)}\hat{\lambda}_0 \to b_0^0$ 代替 $n^{1/(q1+5)}\hat{h}_0 \to a_0^0$，其中 b_0^0 的定义方式类似于 $b_s^0(s=1,\cdots,r)$。定理5.3也需要一点修改（见习题5.1）。

定理5.2和定理5.3是相当强有力的结论，特别是当数据包含离散和连续数据的混合时，有潜力扩展到非参数方法。在5.5节中，通过几个实证应用，我们将说明“不相关”变量经常出现于各种数据集，并且在样本外预测方面，交错鉴定条件密度估计可以胜过一些广泛使用的参数方法，甚至对离散单元的数量与样本大小相当的情形也是如此。

5.4　多元因变量的情形

在这一节我们将考虑当 Y 也是一个一般的多元向量时，给定 X，Y 的条件概率密度函数的估计。令 $Z=(X,Y)$，我们也可写成 $Z=(Z^c,Z^d)$，其中 Z^d 由 r 个离散变量组成，其中 $Z^c\in\mathbb{R}^q$ 代表连续元素。为了表达的简洁，我们首先考虑 Z^d 是一个没有自然的顺序的名义变量向量的情形，$Z^d\in\prod_{s=1}^{r}\{0,1,\cdots,c_s-1\}$。我们记 $Y=(Y^c,Y^d)$，$X=(X^c,X^d)$，并假定 Y^c 包含 Z^c 的前 q_y 个连续元素，而 Y^d 包含 Z^d 的前 r_y 个离散元素。因此，$Y^c\in\mathbb{R}^{q_y}$，$Y^d\in\prod_{s=1}^{r_y}\{0,1,\cdots,c_s-1\}$，$X^c\in\mathbb{R}^{q-q_y}$。类似地，$X^d\in\prod_{s=r_y+1}^{r}\{0,1,\cdots,c_s-1\}$。

与以前一样，令 $f(z)=f(y,x)$ 表示 (Y,X) 的联合概率密度函数。[①] $\mu(x)$ 表示 X 的边际概率密度函数，$g(y|x)=f(y,x)/\mu(x)$ 在给定 $X=x$ 下表示 Y 的条件概率密度函数。

我们用 Z_{is}^d 表示 Z_i^d 的第 s 个元素。同5.1节一样，对于 $Z_{is}^d,Z_{js}^d\in\{0,1,\cdots,c_s-1\}$，我们定义一个一元核函数：如果 $Z_{is}^d=Z_{is}^d$，$l(Z_{is}^d,Z_{js}^d,\lambda_s)=1-\lambda_s$；如果 $Z_{is}^d\neq Z_{is}^d$，$l(Z_{is}^d,Z_{js}^d,\lambda_s)=\lambda_s/(c_s-1)$。乘积核由 $L_{\lambda,Z_i^d,Z_j^d}=\prod_{s=1}^{r}l(Z_{is}^d,Z_{js}^d,\lambda_s)$ 给出。

165

令 Z_{is}^c 表示 Z_i^c 的第 s 个元素，$w(\cdot)$ 表示一个一元核函数，$W(\cdot)$ 表示一个关于 Z^c 的乘积核函数，我们记 $W_{h,Z_i^c,Z_j^c}\overset{\text{def}}{=}\prod_{s=1}^{q}h_s^{-1}w((Z_{is}^c-Z_{js}^c)/h_s)$。

---

① 根据情况，我们有时候把联合概率密度函数记为 $f(z^c,z^d)$ 而不是 $f(y,x)$，其中 $z^c=(y^c,x^c)$，$z^d=(y^d,x^d)$。

为了避免引入太多的符号,我们将使用相同的符号 $L(\cdot)$ 和 $W(\cdot)$ 来表示 Y^d 和 Y^c 的乘积核,即 $L_{\lambda_y,Y_i^d,Y_j^d}=\prod_{s=1}^{r_y}l(Y_{is}^d,Y_{js}^d,\lambda_s)$,$W_{h_y,Y_i^c,Y_j^c}=\prod_{s=1}^{q_y}h_s^{-1}w((Y_{is}^c-Y_{js}^c)/h_s)$。类似地,我们定义 $L_{\lambda_x,X_i^d,X_j^d}=\prod_{s=1}^{r_x}l(X_{is}^d,X_{js}^d,\lambda_{r_y+s})$,$W_{h_x,X_i^c,X_j^c}\overset{\text{def}}{=}\prod_{s=1}^{q-q_y}h_{q_y+s}^{-1}((X_{is}^c-X_{js}^c)/h_{q_y+s})$。

通过下式估计 $f(z)$ 和 $\mu(x)$:

$$\hat{f}(y,x)=\hat{f}(z)=\frac{1}{n}\sum_{i=1}^{n}K_{\gamma Z_i,z},\hat{\mu}(x)=\frac{1}{n}\sum_{i=1}^{n}K_{\gamma_x,X_i,x} \tag{5.18}$$

其中 $K_{\gamma,Z_i,z}=L_{\lambda,Z_i^d,z^d}W_{h,Z_i^c,z^c}$,$K_{\gamma_x,X_i,x}=L_{\lambda_x,X_i^d,x^d}W_{h_x,X_i^c,x^c}$。因此,我们通过下式估计 $g(y|x)=f(x,y)/\mu(x)$:

$$\hat{g}(y\mid x)=\frac{\hat{f}(x,y)}{\hat{\mu}(x)} \tag{5.19}$$

我们通过交错鉴定法来选择平滑参数,它最小化一个样本的加权积分均方误差 I_n,其中

$$\left(\int dz=\sum_{z^d}\int dz^c\right)I_n=\int[\hat{g}(y\mid x)-g(y\mid x)]^2\mu(x)dz=I_{1n}-2I_{2n}+I_{3n}$$

$$I_{1n}=\int[\hat{g}(y\mid x)]^2\mu(x)dz$$

$$I_{2n}=\int\hat{g}(y\mid x)g(y\mid x)\mu(x)dz$$

$$I_{3n}=\int[g(y\mid x)]^2\mu(x)dz$$

同前文一样,I_{3n} 独立于 (h,λ)。因此,关于 (h,λ) 最小化 I_n 等价于最小化 $I_{1n}-2I_{2n}$。

定义 $\hat{f}(X_l,Y_l)$ 和 $\hat{\mu}(X_l)$ 的去一估计量为:

$$\hat{f}_{-l}(X_l,Y_l)=\frac{1}{n}\sum_{i=1,i\neq l}^{n}K_{Z_i,Z_l}$$

$$\hat{\mu}_{-l}(X_l)=\frac{1}{n}\sum_{i=1,i\neq l}^{n}K_{X_i,X_l} \tag{5.20}$$

166 同样地,定义 $G_{-l}(X_l)$ 为

$$\hat{G}_{-l}(X_l)=n^{-2}\sum_{i\neq l}\sum_{j\neq l}K_{X_i,X_l}K_{X_j,X_l}K_{Y_i,Y_j}^{(2)} \tag{5.21}$$

其中 $K_{Y_i,Y_j}^{(2)}=\sum_{y^d}\int K_{\gamma_y,Y_i,y}dy^c$。

然后,通过使用了 5.1 节中类似的论证,Racine 等(2004)证明了 $I_{1n}-I_{2n}$ 的

一个一致估计量：

$$\mathrm{CV}(h,\lambda) \overset{\text{def}}{=} \frac{1}{n}\sum_{l=1}^{n} \frac{\hat{G}_{-l}(X_l)}{[\hat{\mu}_{-l}(X_l)]^2} - \frac{2}{n}\sum_{l=1}^{n} \frac{\hat{f}_{-l}(X_l,Y_l)}{\hat{\mu}_{-l}(X_l)} \tag{5.22}$$

其中$\hat{f}_{-l}(X_l,Y_l)$,$\hat{\mu}_{-l}(X_l)$,$\hat{G}_{-l}(X_l)$的定义在式(5.20)和式(5.21)中给出。

接着，我们选择$(h,\lambda)=(h_1,\cdots,h_q,\lambda_1,\cdots,\lambda_r)$去最小化式(5.22)中定义的目标函数CV$(h,\lambda)$。用$(\hat{h},\hat{\lambda})$表示交错鉴定的$(h,\lambda)$的选择。由于我们已经知道无关的独立变量将被渐近地平滑掉，因此，下面的分析我们只考虑所有独立变量都相关的情况。Racine 等(2004)证明了$\hat{h}_s/h_s^0 \xrightarrow{p} 1$，$\hat{\lambda}_s/\lambda_s^0 \xrightarrow{p} 1$，其中

$$h_s^0 = c_{1s}n^{-1/(4+q)}, s = 1,\cdots,q$$
$$\lambda_s^0 = c_{2s}n^{-2/(4+q)}, s = 1,\cdots,r$$

其中c_{1s}和c_{2s}是常数，(h^0,λ^0)是最小化交错鉴定函数首项的非随机最优平滑参数 (5.23)

定理5.4得出了$(\hat{h},\hat{\lambda})$趋于(h^0,λ^0)的收敛速度，证明见 Racine 等(2004)。

定理 5.4 在 Racine 等(2004)给出的假定下，我们有$(\hat{h}_s-h_s^0)/h_s^0 = O_p(n^{-\alpha/(4+q)})$，$s=1,\cdots,q$，以及$\hat{\lambda}_s-\hat{\lambda}_s^0=O_p(n^{-\beta})$，$s=1,\cdots,r$，其中$\alpha=\min\{2,q/2\}$，$\beta=\min\{1/2,4/(4+q)\}$。

给定定理5.4，我们可以进一步证明以下定理：

定理 5.5 定义

$$B_{1s}(z) = (1/2)\kappa_2 f_{ss}(y \mid x)/\mu(x), s = 1,\cdots,q_y$$

令

$$B_{1s}(z) = (1/2)\kappa_2[f_{ss}(z) - \mu_{ss}(x)g(y \mid x)]/\mu(x), s = q_y+1,\cdots,q$$

同样地，定义

$$B_{2s} = \frac{1}{c_s-1}\sum_{v^d\in\mathcal{D}} I_s(v^d,z^d)f(z^c,v^d), s = 1,\cdots,r_y$$

$$B_{2s}(z) = \frac{1}{c_s-1}\sum_{u^d\in\mathcal{D}_x} I_s(u^d,x^d)[f(z^c,y^d,u^d) - g(y \mid x)\mu(x^c,u^d)]/\mu(x), s = r_y+1,\cdots,r$$

又因为$\Omega(z)=\kappa^q g(y|x)/\mu(x)$，所以

$$\sqrt{nh_1\cdots h_q}\left[\hat{g}(y \mid x) - g(y \mid x) - \sum_{s=1}^{q}\hat{h}^2 B_{1s}(z) - \sum_{s=1}^{r}\hat{\lambda}_s B_{2s}(z)\right]$$

依分布收敛于$N(0,\Omega(z))$。

定理5.5的证明在5.4.2节给出。

我们现在考虑怎样把上面的分析延伸到Z^d包含有序的分类变量的情形。

5.4.1 广义分类数据的情形

关于有序的分类变量, Aitchison 和 Aitken(1976, p. 29)建议使用加权核函数:

$$l_{AA}(Y_{is}^d, Y_{js}^d, \lambda_s) = \begin{bmatrix} c_s \\ t \end{bmatrix} \lambda_s^t (1-\lambda_s)^{c_s-t}, \text{当} \mid Y_{is}^d - Y_{js}^d \mid = t \text{时}, 0 \leqslant t \leqslant c_s \tag{5.24}$$

其中$\begin{bmatrix} c_s \\ t \end{bmatrix} = c_s!/[t!(c_s-t)!]$。正如先前所讨论的,这个核函数有一个优点,那就是它导致了一个合适的密度估计量(核 l_{AA}加总为1),但是它有一个缺点,就是不能平滑掉无关变量。因为无关的离散变量仅仅出现在协变量 X^d 中,我们建议仅仅对于离散有序因变量 Y^d,使用在式(5.24)中定义的核 l_{AA},而对于有序离散自变量 X^d,我们建议使用下面的核函数:

$$l(X_{s,i}^d, X_{s,j}^d, \lambda_s) = \lambda_s^t, \text{当} \mid X_{s,i}^d - X_{s,j}^d \mid = t \text{时}, 0 \leqslant t \leqslant c_s \tag{5.25}$$

168 其中 $\lambda_s \in [0,1]$。当 $\lambda_s = 1$ 时,我们得到一个均匀的权重1,所以在核函数这种选择下,X^d 中的无关元素将被自动移除。注意,我们并不需要关于 X^d 的核加总为1的条件,因为在条件密度估计中,关于 X^d 的核同时出现在分子和分母之中,因此,任何非零常数乘以这个核函数都将在分子和分母中被消掉,使得估计结果不变。

定理5.4和定理5.5的结果同样可以容易地被推广到某些离散变量具有自然顺序的情形,除了当 x_s^d 具有自然顺序时,$\frac{1}{c_s-1}\mathbf{1}_s(u^d, x^d)$ 必须用 $\mathbf{1}(\mid u_s^d - x_s^d \mid = 1)\prod_{t \neq s}\mathbf{1}(u_t^d = x_t^d)$ 替代,当 y_s^d 具有自然顺序时(因为$\begin{bmatrix} c_s \\ 1 \end{bmatrix} = 1/c_s$),$\frac{1}{c_s-1}\mathbf{1}_s(v^d, y^d)$ 必须用$\frac{1}{c_s}\mathbf{1}(\mid y_s^d - v_s^d \mid = 1)\prod_{t \neq s}\mathbf{1}(y_t^d = v_t^d)$ 替代。

5.4.2 定理5.5的证明

证明:通过使用随机等连续(stochastic equicontinuity)理论,我们知道无论我们使用随机平滑参数$\hat{h}_s, \hat{\lambda}_s$ 还是使用式(5.23)中定义的非随机平滑参数 $h_s = c_{1s}n^{-1/(4+q)}$和 $\lambda_s = c_{2s}n^{-2/(4+q)}$, $\hat{g}(y|x)$的渐近分布保持不变。因此,我们在证明中只考虑非随机平滑参数的情形。

令

$$K_{\gamma_y, y_i, y} = W_{h_y, y_i^c, y^c} L_{\lambda_y, y_i^d, y^d}, K_{\gamma_x, y_i, y} = W_{h_x, x_i^c, x^c} L_{\lambda_x, x_i^d, x^d}$$

我们分别估计 $f(y,x)$、$\mu(x)$ 和 $g(y|x)$：

$$\hat{f}(y,x) = \frac{1}{n}\sum_{i=1}^{n} K_{\gamma_y,y_i,y} K_{\gamma_x,x_i,x}$$

$$\hat{\mu}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{\gamma_x,x_i,x}$$

$$\hat{g}(y \mid x) = \frac{\hat{f}(y,x)}{\hat{\mu}(x)}$$

为了得到 $\hat{g}(y|x) - g(y|x)$ 的渐近正态性，我们记

$$\hat{g}(y \mid x) - g(y \mid x) = \frac{[\hat{g}(y \mid x) - g(y \mid x)]\hat{\mu}(x)}{\hat{\mu}(x)} \equiv \frac{\hat{m}(y,x)}{\hat{\mu}(x)}$$

其中 $\hat{m}(y,x) = [\hat{g}(y|x) - g(y|x)]\hat{\mu}(x) \equiv \hat{f}(y,x) - g(y|x)\hat{\mu}(x)$。

我们接下去计算 $E[\hat{m}(y,x)]$ 和 $\mathrm{var}(\hat{m}(y,x))$。注意

$$E[\hat{m}(y,x)] = E[\hat{f}(y,x)] - g(y \mid x)E[\hat{\mu}(x)] \tag{5.26}$$

而

$$E[\hat{f}(y,x)] = f(y,x) + \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2 f_{ss}(y,x) + \sum_{s=1}^{r}\frac{\lambda_s}{c_s - 1}\sum_{v^d \in \mathcal{D}} I_s(v^d,z^d) f(z^c,v^d) + o(|h|^2 + |\lambda|) \tag{5.27}$$

其中 $f_{ss}(z) = \partial f^2(z)/\partial(z_s^c)^2 (s = 1,\cdots,q)$，$\kappa_2 = \int k(v)^2 \mathrm{d}v$，$|h|^2 = \sum_{s=1}^{q} h_s^2$，$|\lambda| = \sum_{s=1}^{q}\lambda_s$。

类似地，

$$E[\hat{\mu}(x)] = \mu(x) + \frac{\kappa_2}{2}\sum_{s=q_y+1}^{q} h_s^2 \mu_{ss}(x) + \sum_{s=r_y+1}^{r}\frac{\lambda_s}{c_s - 1}\sum_{u^d \in \mathcal{D}_x} I_s(u^d,x^d)\mu(x^c,u^d) + o(|h_x|^2 + |\lambda_x|) \tag{5.28}$$

其中 $|h_x|^2 = \sum_{s=q_y+1}^{q} h_s^2$，$|\lambda_x| = \sum_{s=r_x+1}^{r}\lambda_s$。

将式(5.27)和式(5.28)代入式(5.26)，我们得到

$$E[\hat{m}(y,x)] = \sum_{s=1}^{q} h_s^2 B_{1s}(z) + \sum_{s=1}^{r}\lambda_s B_{2s}(z) + o(|h|^2 + |\lambda|) \tag{5.29}$$

170　接着，记 $H_q = h_1 \cdots h_q$，$\int dz = \sum_{z^d} \int dz^c$，我们有

$$\begin{aligned}
\operatorname{var}[\hat{m}(y,x)] &= \operatorname{var}[\hat{f}(y,x) - g(y \mid x)\hat{\mu}(x)] \\
&= \frac{1}{n}\{\operatorname{var}([K_{\gamma_y,y_i,y} - g(y \mid x)]\kappa_{\gamma_x,x_i,x})\} \\
&= \frac{1}{n}\{E\{[K_{\gamma_y,y_i,y} - g(y \mid x)]^2 K^2_{x_i,x}\} + O(1)\} \\
&= \frac{1}{n}\{E[K^2_{\gamma_y,y_i,y} K^2_{\gamma_x,x_i,x}] + O(1)\} \\
&= n^{-1}\left\{\int f(z_i) K^2_{\gamma,z_i,z} dz_i + O(1)\right\} \\
&= n^{-1}\left\{H_q^{-1}\int f(z^c + hv, z^d) W^2(v) dv + o(H_q^{-1})\right\} \\
&= (nH_q)^{-1}\kappa^q f(y,x) + (\text{s.o.})
\end{aligned} \tag{5.30}$$

式(5.29)和式(5.30)给出了 $\hat{m}(y,x)$ 的偏误首项和方差首项。为了最优平滑，即 $h_s \sim n^{-1/(4+q)}$ $(s=1,\cdots,q)$，$\lambda_s \sim n^{-2/(4+q)}$ $(s=1,\cdots,r)$，那么根据式(5.29)、式(5.30)，并应用 Liapunov 中心极限定理，我们知道

$$\sqrt{nH_q}\left\{\hat{m}(y,x) - \left[\sum_{s=1}^{q} h_s^2 B_{1s}(z) + \sum_{s=1}^{r} \lambda_s B_{2s}(z)\right]\right\} \xrightarrow{d} N(0, \kappa^q f(y,x)) \tag{5.31}$$

注意到这可以得出

$$\hat{\mu}(x) - \mu(x) = O_p\left(\sum_{s=g_y+1}^{q} h_s^2 + (nh_{q_y+1}\cdots h_q)^{-1/2}\right) \tag{5.32}$$

171　联立式(5.31)和式(5.32)，我们立即有：

$$\begin{aligned}
&\sqrt{nH_q}\left[\hat{g}(y \mid x) - g(y \mid x) - \sum_{s=1}^{q} h_s^2 B_{1s}(z) - \sum_{s=1}^{r} \lambda_s B_{2s}(z)\right] \\
&\equiv \sqrt{nH_q}\left\{\hat{m}(y,x) - \hat{\mu}(x)\left[\sum_{s=1}^{q} h_s^2 B_{1s}(z) + \sum_{s=1}^{r} B_{2s}(z)\right]\right\}\Big/\hat{\mu}(x) \\
&= \sqrt{nH_q}\left\{\hat{m}(y,x) - \mu(x)\left[\sum_{s=1}^{q} h_s^2 B_{1s}(z) + \sum_{s=1}^{r} B_{2s}(z)\right]\right\}\Big/\mu(x) + o_p(1) \\
&\xrightarrow{d} \frac{1}{\mu(x)} N(0, \kappa^q f(y,x)) = N(0, \kappa^q g(y \mid x)/\mu(x))
\end{aligned} \tag{5.33}$$

□

5.5 应　用

5.5.1 一个腐败的非参数分析

政治腐败(或相关概念)在政治理论和经济发展理论中发挥着重要作用。不同国家随时间在腐败上是否收敛,或者腐败随时间的持续在分布上是否类似,仍然是一个继续有争议的话题。

腐败感知指数(corruption perception index,CPI)[①]按可感知的腐败情况对各国进行排序。CPI 的取值在 0—10 之间,10 表示不存在腐败。参考 McAdam 和 Rummel(2004)中采用的分析思路,他们考察了一个 40 个国家的面板,包含 1995—2002 年的全部记录,我们创建了一个 45 个国家的平衡面板,包含 1996—2004 年这 9 年的全部记录。对这个有 $n=405$ 个观测值的面板,我们使用本章讨论过的方法,以年份为条件估计出了 CPI 的条件概率密度函数。这一估计的条件概率密度函数在图 5.1 中绘出。

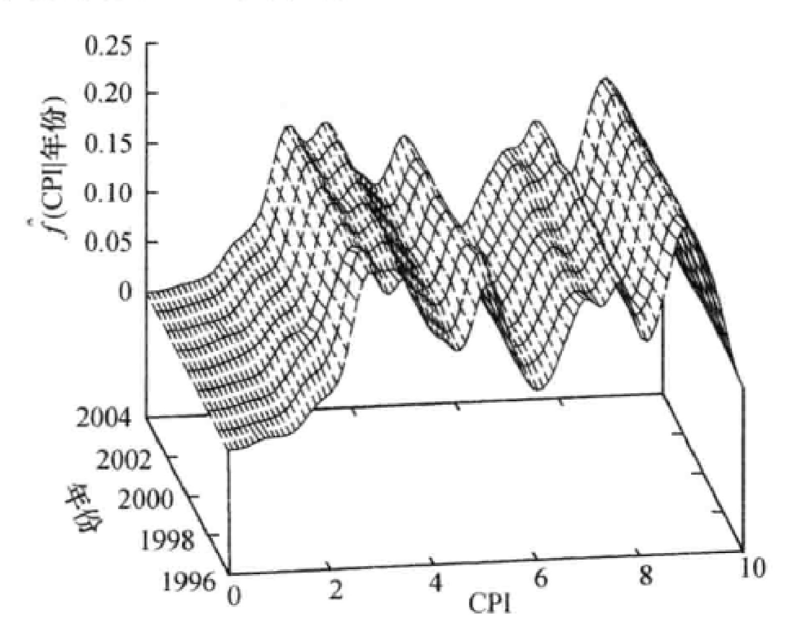


图 5.1　腐败感知指数的条件概率密度函数,1996—2004 年,$n=405$

我们把年份看成有序分类变量,最小二乘交错鉴定法的窗宽是 $\hat{h}_{\text{CPI}}=0.224$ 和 $\hat{\lambda}_{\text{year}}=1.00$($\lambda_{\text{year}}$的上界值)。$\hat{\lambda}_{\text{year}}$的交错鉴定值表明全部年份数据的合并是合理的,即关于时间在分布上没有明显的变化。这个发现支持了持续性的假定(即分布在时间上稳定),而估计出的概率密度函数的形状与多重均衡(即多种模式)一致。这些结果与 McAdam 和 Rummel (2004,p. 509)的一致,他们称"我

172

① http://www.transparency.org/cpi/

们的发现支持理论文献中提出的关注，即腐败可以具有高度的持久性，并具有多种均衡的特征”。

5.5.2 婚外情数据

在一篇被广泛引用的文献中，Fair(1978)提出了一种“婚外情理论”并考虑了个人时间在工作和两种休闲活动即与妻子和情人在一起的时间的分配。这个独特的数据集来自两个杂志的调查，Fair 采用一个参数 Tobit 估计量来对每年的婚外情数据建模。这个数据集以及计量经济学的方法继续引起了学者们的兴趣。Pagan 和 Vella(1989)，Chernozhukov 和 Hong(2002)，Wells(2003)，以及 Li 和 Racine(2004b)对原始研究中的证据和使用的经济计量方法进行了复查。

173

这个研究一个相当吸引人的方面是一个人的个人特征对参加婚外情活动倾向的潜在影响。特别地，Fair(1978)发现夫妻间的不忠实随结婚年数的增加而显著增加。Pagan 和 Vella(1989)通过各种诊断检验对这个发现的有效性提出质疑，并且也提出一个离散计数模型要比 Fair 使用的 Tobit 模型更加合适，而 Wells(2003)所做的分析认为，结婚时间越长人们也许更倾向于“鬼混”。

数据由 9 个变量的 601 个观测值组成，性别(0/1)，年龄(9 个组)，结婚年数(8 个组)，孩子(0/1)，宗教程度(1—5)，受教育水平(7 个组)，职业(7 个组)，婚姻评价(1—5)，在过去的一年中婚外性行为的次数(0 = 没有，1 = 1 次，2 = 2 次，3 = 3 次，7 = 4—10 次，12 = 每月一次或者更加频繁)。显然，这些变量每一个都是分类的。

Fair(1978)所使用的 Tobit 设定中有一个与结婚年数相关联的正的并且显著的参数，这一点与一个 Poisson 计数模型和一个简单线性模型是相同的。然而，这些模型有一个相同的特征就是线性指数设定。由于缺少交互项，一些变量的离散和无序的性质会带来对这些设定合适性的质疑。

自动窗宽选择方法如交错鉴定法的一个特征是通过给相关窗宽选择一个较大的值消除不相关的变量的能力。因此我们应用交错鉴定法条件密度估计量并检验交错鉴定窗宽的表现。我们把性别、孩子、职业看成是无序的，剩下的变量则看成是有序分类变量。

从表 5.1 中我们可以观察到与结婚年数相关的交错鉴定平滑参数位于它们的上界值，这表明结婚年数与婚外情次数的预测是不相关的。实际上，从预测的角度，表现出相关的变量只有年龄、被访者声称的宗教程度，以及不意外地，被访者对其婚姻的评价。

174

表 5.1　条件密度,最小二乘交错鉴定法

变量	$\hat{\lambda}$	上界
婚外情次数	0.019	0.833
性别	0.500	0.500
孩子	0.500	0.500
职业	0.857	0.857
年龄	0.886	1.000
结婚年数	1.000	1.000
宗教	0.290	1.000
教育	1.000	1.000
婚姻评价	0.329	1.000

为了更好地理解为什么参数和非参数方法产生不同的结果,我们更加详细地讨论非参数的估计结果。图 5.2 画出了非参数模型中当所有其他的协变量固定在它们的中位数不变时,以年龄为条件的婚外情发生的期望次数。我们观察到年龄和婚外情次数之间的关系是非线性的,较年轻时是平的和/或向上倾斜的,对年龄超过 40 的人们表现出一个向下的趋势。因此,非参数的结果表明年龄对婚外情次数有一个非线性的影响,然而,在年龄和其他相关的协变量条件下,结婚年数与婚外情次数是独立的。结婚年数和年龄是高度相关的协向量,并且简单参数模型假定每个协变量要么只有一个全局为正的影响,要么仅有一个负的影响。当真实关系是非线性时,比如年龄,在一个范围内年龄有正的影响,而在另一个范围内则有负的影响。参数模型表现出错误地认为年龄协变量有一个负效应,并试图用线性指数函数为年龄中的非线性建模(更年轻时向上倾斜的部分),参数模型引入一个不相关协变量(其与年轻的年龄高度相关)并且为

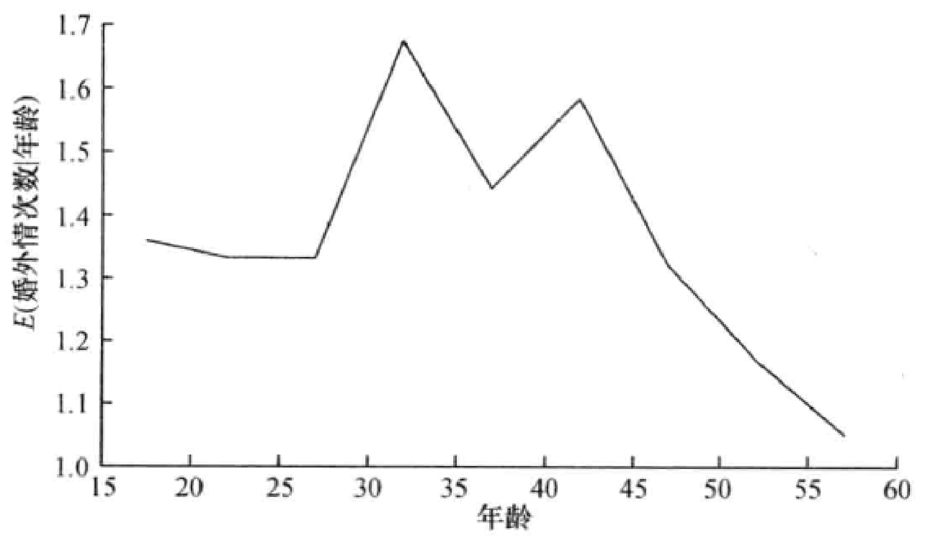


图 5.2　在年龄条件下的期望婚外情次数,保持变量在它们的中位数不变

它分配了一个显著的正的符号。

上面的分析是基于这样的假定,非参数方法对数据比参数设定更加可靠。175 为了评估是否的确是这样的情况,我们对核方法和表现最好的参数模型(有序 Logit 设定)考虑了样本内预测。非参数方法和有序 Logit 正确预测婚外情次数分别达到 79.2% 和 75.0% 。同样,对于 150 个报告有正婚外情次数的人来说,非参数方法正确预测的次数是 25 人(16.7%),有序 Logic 模型正确预测的只有 3 人(2.0%)。预测结果表明参数模型是错误设定的。

Fair(1978)中所研究的婚外情数据是一个继续产生争论的丰富的数据集。现有的工作主要关注基于参数线性单一指数模型的设定检验。相关协变量可能的非线性在很大程度上被忽视了。我们转而关注参数错误设定对评价预测值的影响,并得出结论,一个长期被认为是一个显著的婚外情预测元的协变量,结婚年数,实际上没有表现出与一个人参与婚外情活动增加的倾向是关联的。

5.5.3 已婚女性劳动参与率

176 我们现在来考虑一个简单的应用,我们为已婚瑞士女性参与劳动的决定建模,只以她们的非劳动收入为基础。我们使用 Gerfin(1996)的数据,文中使用了一个样本大小为 $n=872$ 的横截面数据集为瑞士已婚女性的劳动市场参与率建模。为了当前的目的,我们简单考虑使用核估计来得到参与率的条件概率,它是非劳动收入的一个函数。我们还采用了一个被 Gerfin 使用的含有线性指数的 Probit 模型。关于核方法,Aitchison 和 Aitken(1976)核被用于二元的劳动参与率变量,高斯核被用于连续的非劳动收入的变量,窗宽由极大似然交错鉴定法选择($\lambda_{\text{lfp}}=0.07, h_{\text{nonlaborinc}}=0.16$)。为了防止由似然法所造成的潜在的过度平滑,我们也计算了最小二乘交错鉴定窗宽($\lambda=0.03, h=0.11$),以及一个特别的值 $h=x_{sd}n^{-1/5}=0.11$,其中 x_{sd} 是非劳动收入对数的样本标准差。不同的方法得出相似的 h 值,支持了在这个应用中似然交错鉴定法的使用(即用各种方法去选择 λ 和 h,人们得到相似的结果)。

首先,在图 5.3 中我们注意到参数和非参数方法大体上是一致的,表明了一种可能性,随着非劳动收入的增加劳动参与率从超过 50% 下降到 20% 左右。177 然而,核方法发现了在低水平的非劳动收入下劳动参与率有一个上升,然后随着非劳动收入的提高开始下降。Juhn 和 Murphy(1997)对此提出了一个可能的解释,他们认为低工作能力的女性倾向于嫁给低工作能力的男性;并且低工作能力的男性收入也低,这意味着他们妻子的非劳动收入也倾向于较低。对于低水平的非劳动收入,这些女性因其低观测值(即能力),她们在家里的生产力高于在劳动市场的生产力。然而,随着男性收入的增加,女性的观测值(能力)倾

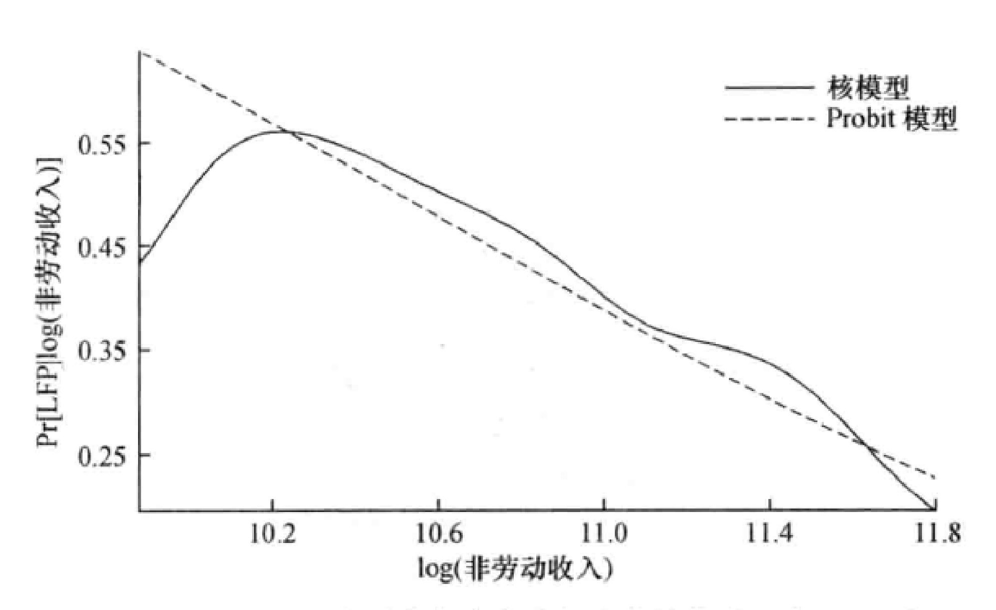


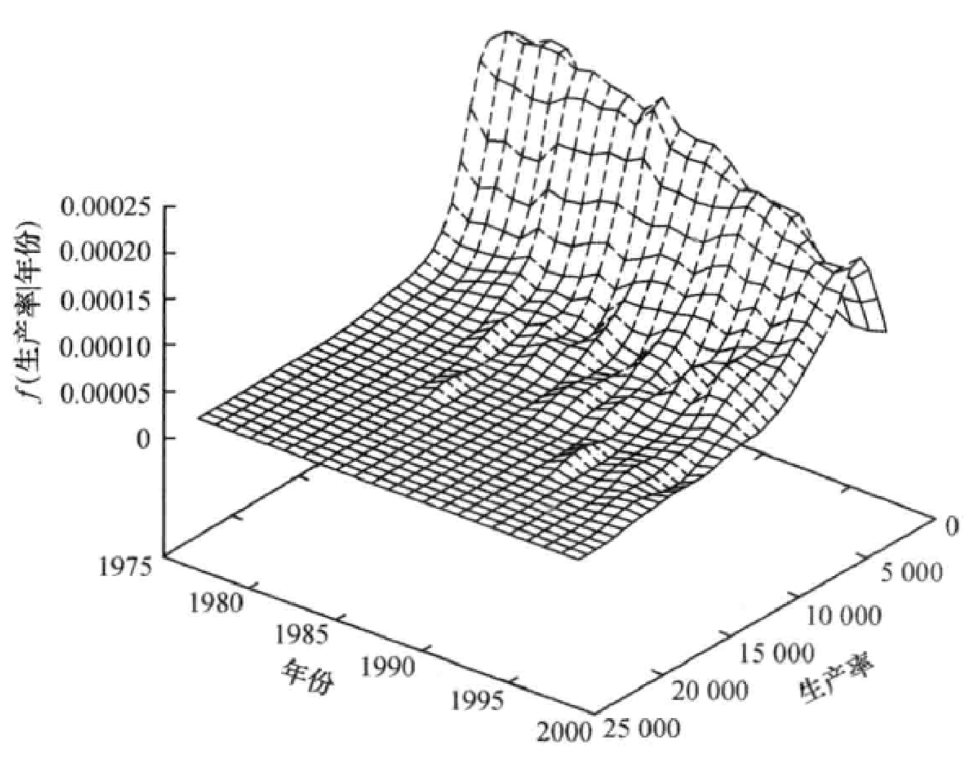
图 5.3　Gerfin 数据劳动参与率条件概率的核估计($P[Y=1|x]$)

向于提高,这就导致更高的女性劳动参与率。在某些时候,当非劳动收入的影响(即财富效应)开始按预期占优势时,女性劳动参与率就开始下降。

5.5.4　劳动生产率

我们使用一个在 van Dijk 和 Szirmai(2003)中报告的非平衡面板数据来衡量印度尼西亚纸浆和造纸公司在 1975—1997 年的劳动生产率。这个数据集有一些有趣的特征。第一个印度尼西亚造纸工厂建立于 1923 年,直到 1974 年才仅仅建立了 6 个国有工厂。印刷和书写纸市场在 1974—1984 年间发展迅速,工厂的数量也从 7 个上升到了 33 个,并且所有新建工厂都是私人所有的。在这个面板中,一些年份(单元)非常稀疏,少到只包含 8 个观测数据,而最大的单元仅包含 52 个观测数据,总共有 $n=783$ 个观测值。因此在离散协变量年份的条件下,我们考虑为劳动生产率的条件概率密度函数建模。高斯核用于劳动生产率,Aitchison 和 Aitken(1976)核被用于年份。最大似然交错鉴定被用于窗宽选择,其中 $h_{\text{labor}}=1\,044.07$,$\lambda_{\text{year}}=0.28$。

在 1983 年引入的一个主要的政策改变是旨在鼓励这个行业的增长,因此观察在 1983 年前后生产率是否表现出不同是有意义的。在图 5.4 中,我们观察到劳动生产率在 20 世纪 80 年代末直到整个 90 年代有一个变动。当单元的数量相对于样本大小较大时,这个例子突出了来自平滑离散数据带来的潜在好处。



178 图 5.4 在时间条件下劳动生产率概率密度函数的核估计

5.5.5 多元变量 Y 条件密度例子:OECD 条件下的 GDP 增长和人口增长

我们采用 Maasoumi, Racine 和 Stengos(2007)所使用的数据集,他们考察了有关经济增长速度和存在"收敛俱乐部"的问题。对于目前的例子,我们的目标只是说明用窗宽选择的数据驱动方法去估计多元 Y 条件概率密度函数是直接的。用例子中的方法,我们把 88 个国家的 7 个五年期的面板数据合并,分别是 1960—1964 年,1965—1969 年,1970—1974 年,1975—1979 年,1980—1984 年,1985—1989 年,1990—1994 年,总共有 $n=616$ 个观测数据。我们考虑以下的变量:y_1,每个时期的人均收入增长率;y_2,在每个时期每年人口的增长率;x, OECD 身份(0/1)。人均 GDP 的年平均增长率和每期的人口数量来自世界银行。我们估计条件密度 $f(y_1,y_2|x)$并在图 5.5 中给出条件概率密度函数的结果,也就是绘制 $\hat{f}(y_1,y_2|X=0)$和 $\hat{f}(y_1,y_2|X=1)$。Hall 等(2004)采用最小二乘交错鉴

179

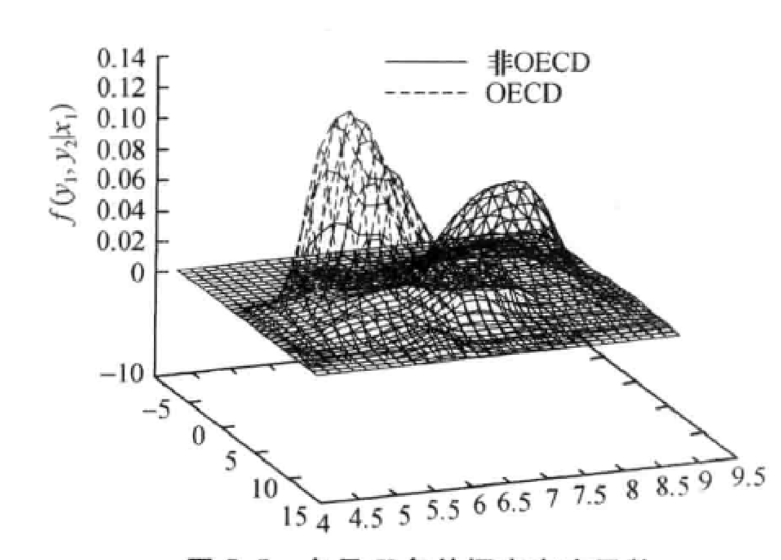


图 5.5 多元 Y 条件概率密度函数

定法得到，其中$\hat{h}_{\dot{GDP}}=0.96$，$\hat{h}_{\dot{POP}}=0.21$，$\hat{\lambda}_{OECD}=0.00$。

图5.5给出了多元Y条件概率密度函数的清晰画面。非OECD国家的联合概率密度函数有约7.3%的人口增长率模式，然而OECD国家在此期间约为5.7%，非OECD国家的GDP增长率模式约为1.8%，OECD国家约为2.7%。

5.6 习 题

180

习题5.1 当Y为一个离散变量时，比如，

$$Y \in S_0^d = \{0,1,\cdots,c_0-1\}$$

证明式(5.17)应该被修改为：

$$(n\hat{h}_1\cdots\hat{h}_{q_1})^{1/2}\left(\bar{g}(y\mid x)-g(y\mid x)-\sum_{s=1}^{q_1}B_{1s}(\bar{x},y)\hat{h}_s^2-\sum_{s=1}^{r_1}B_{2s}(\bar{x},y)\hat{\lambda}_s\right)$$

$$\to N(0,\sigma_g^2(\bar{x},y))$$

其中

$$B_{1s}(\bar{x},y)=\frac{1}{2}\kappa_2\sum_{s=1}^{q_1}\hat{h}_s^2\left\{\frac{\bar{f}_{ss}(y\mid\bar{x})}{\bar{\mu}(\bar{x})}-\frac{\bar{\mu}_{ss}(\bar{x})}{\bar{\mu}(\bar{x})}\bar{g}(y\mid\bar{x})\right\}$$

$$B_{2s}(\bar{x},y)=\frac{\hat{\lambda}_0}{c_0-1}\sum_{Z^d\in S_0^d}\mathbf{1}_0(z^d,y)\left\{\bar{g}(z\mid\bar{x}^c,\bar{v}^d)-\frac{\bar{\mu}(\bar{x}^c,\bar{v}^d)}{\bar{\mu}(\bar{x})}\bar{g}(y\mid\bar{x})\right\}$$

$$+\sum_{s=1}^{r_1}\frac{\hat{\lambda}_s}{c_s-1}\sum_{\bar{v}^d}\mathbf{1}_s(\bar{v}^d,x_s^d)\left\{\bar{g}(y\mid\bar{x}^c,\bar{v}^d)-\frac{\bar{\mu}(\bar{x}^c,\bar{v}^d)}{\bar{\mu}(\bar{x})}\bar{g}(y\mid\bar{x})\right\}$$

$$\sigma_g^2(\bar{x},y)=k^{q_1}\bar{g}(y\mid\bar{x})/\bar{\mu}(\bar{x})$$

其中$\hat{\lambda}_0$是用交错鉴定法选择的关于y的平滑参数。

习题5.2 运用我们已经在习题2.11中使用过的，由Pagan和Ullah(1999, pp.154—155)给出的数据集，运用一个收入组合的局部常数核估计(对不同年龄的收入取对数)，用最小二乘交错鉴定法生成以年龄为条件的收入的概率密度函数。

接着，画出得到的条件概率密度函数，对比在习题2.11中用条件均值函数得到的估计。你能很好地想象从条件密度函数中得到的条件均值函数吗？

第6章 条件累积分布函数与分位数估计

181

我们在第2章讨论的条件均值函数(即回归函数)毫无疑问在所有的应用数据分析中是最流行的。然而条件均值有时不是人们可能考虑的最适用的目标。比如,如果因变量是截断的,基于这个截断数据的条件均值函数可能极大地偏离了和未截断的总体相关联的真实条件均值函数。后者是应用经济学者特别感兴趣的。在这种情况下,所谓的分位数回归函数可能更加稳健,并提供了一种可行的替代条件均值回归函数的选择。除了当截断数据出现时稳健外,还可以通过估计一系列条件分位数得到因变量条件分布的一个更加全面的图像而不仅仅是简单地呈现条件均值本身。对于分位数回归方法的一个完整处理可参见 Koenker(2005)。

在这一章,我们开始研究在非参数背景中估计一个条件累积分布函数。我们首先讨论只有连续协变量的条件累积分布函数的估计。经常让人直接关注的条件分位数函数可以通过条件累积分布函数直接得到,以及通过对条件累积分布函数进行简单的转化得到。在考虑了只有连续数据情况后,我们接着证明这个结论可以容易地扩展到混合了离散和连续协变量的情形,这是应用背景下经常面临的情形。

182

6.1 没有平滑因变量情况下估计具有连续协变量的条件累积分布函数

我们首先考虑一个没有平滑因变量的条件累积分

布函数估计量。这个估计量在思想上与我们在第2章中讨论的局部常数回归估计量的估计式类似。我们使用 $F(y|x)$ 表示在给定 $X=x$ 时 Y 的累积分布函数，$\mu(x)$ 是 X 的边际密度函数。注意 $F(y|x)=E[\mathbf{1}(Y_i\leqslant y)|X_i=x]$ 是在 $X_i=x$ 条件下 $\mathbf{1}(Y_i\leqslant y)$ 的条件均值函数，这说明人们可以使用在第4章中讨论的条件均值（回归）函数估计量去估计 $F(y|x)$。于是，我们估计 $F(y|x)$ 为

$$\tilde{F}(y\mid x)=\frac{n^{-1}\sum_{i=1}^{n}\mathbf{1}(Y_i\leqslant y)W_h(X_i,x)}{\hat{\mu}(x)} \tag{6.1}$$

其中，$\hat{\mu}(x)=n^{-1}\sum_{i=1}^{n}W_h(X_i,x)$ 是 $\mu(x)$ 的核估计量，$W_h(X_i,x)=\prod_{s=1}^{q}h_s^{-1}w((X_{is}-X_s)/h_s)$，且 $w(\cdot)$ 是一个一元核函数。

我们做出如下的假定：

假定6.1 $\mu(x)$ 和 $F(y|x)$ 两者关于 x^c 都具有连续的二阶导数。$w(\cdot)$ 是对称、有界、具有紧支撑的概率密度函数。

假定6.2 假定当 $n\to\infty$ 时，对于所有的 $s=1,\cdots,q$，$nh_1\cdots h_q\to\infty$，$h_s\to 0$。

对于 $s=1,\cdots,q$，令 $A_s(y,x)=\partial A(y,x)/\partial x_s$，$A_{ss}(y,x)=\partial^2A(y,x)/\partial x_s^2$，$A_0(y,x)=\partial A(y,x)/\partial y$，$A_{00}(y,x)=\partial A^2(y,x)/\partial y^2$，$\kappa_2=\int w(v)v^2\mathrm{d}v$，$k=\int w(v)^2\mathrm{d}v$，其中 $\kappa^q=\prod_{s=1}^{q}\int w(v_s)^2\mathrm{d}v_s=\int W(v)^2\mathrm{d}v$。

我们记 $\tilde{F}(y|x)-F(y|x)=\tilde{M}(y,x)/\hat{\mu}(x)$，其中 $\tilde{M}(y,x)=[\tilde{F}(y|x)-F(y|x)]\hat{\mu}(x)$。下面的定理给出了 $\tilde{F}(y|x)$ 的渐近分布。

定理6.1 在假定6.1和6.2下，也假定 $\mu(x)>0$ 和 $F(y|x)>0$，我们有

(i) $E[\tilde{M}(y,x)]=\mu(x)\left[\sum_{s=1}^{q}h_s^2B_s(y,x)\right]+o\left(\sum_{s=1}^{q}h_s^2\right)$，其中

$$B_s(y,x)=(1/2)\kappa_2[F_{ss}(y\mid x)+2\mu_s(x)F_s(y\mid x)/\mu(x)]$$

183

(ii) $\mathrm{var}[\tilde{M}(y,x)]=(nh_1\cdots h_q)^{-1}\mu(x)^2\Sigma_{y|x}+o((nh_1\cdots h_q)^{-1})$，其中

$$\Sigma_{y|x}=\kappa^qF(y\mid x)[1-F(y\mid x)]/\mu(x)$$

(iii) 如果 $(nh_1\cdots h_q)^{1/2}\sum_{s=1}^{q}h_s^3=o(1)$，那么

$$(nh_1\cdots h_q)^{1/2}\left[\tilde{F}(y\mid x)-F(y\mid x)-\sum_{s=1}^{q}h_s^2B_s(y,x)\right]\xrightarrow{d}N(0,\Sigma_{y|x})$$

定理6.1的证明在6.9节给出。

利用

$$\tilde{F}(y \mid x) - F(y \mid x) = \widetilde{M}(y,x)/\hat{\mu}(x)$$
$$= \widetilde{M}(y,x)/\mu(x) + O_p\left(\sum_{s=1}^{q} h_s^2 + (nh_1 \cdots h_q)^{1/2}\right)$$

定理 6.1 意味着

$$\text{MSE} \mid \tilde{F}(y \mid x) \mid = \left[\sum_{s=1}^{q} h_s^2 B_s(y,x)\right]^2 + \frac{\Sigma_{y \mid x}}{nh_1 \cdots h_q} + (\text{s.o.}) \qquad (6.2)$$

其中,(s. o.)表示(被忽略的)更低阶项。

如果我们选择 $h_1, \cdots, h_q$ 最小化一个由 $\int \text{MSE}[\tilde{F}(y \mid x)] s(y,x) \mathrm{d}x\mathrm{d}y$ 给出的加权积分均方误差,这里 $s(y,x)$ 是一个非负的权重函数,那么从式(6.2)中可以容易证明最小化积分均方误差的最优平滑参数应该是 $h_s \sim n^{-1/(4+q)}$ $(s = 1, \cdots, q)$。然而,对最优的 h_s 不存在一个闭式表达式。尽管我们可以计算插入的 h_s,至少可以说它在计算上是有挑战的。Li 和 Racine(2007)建议使用为估计条件 PDF 设计的最小二乘交错鉴定法来选择窗宽。在第 5 章我们已经证明了估计条件概率密度函数时,最优平滑得出的窗宽形式为 $h_0 \sim h_s \sim n^{-1/(5+q)}$(所有 $s = 1, \cdots, q$)。注意指数为 $-1/(5+q)$,因为我们假定 Y 为一个连续变量,(y,x) 的维数是 $q+1$。我们可以将条件最小二乘交错鉴定平滑参数乘以因子 $n^{1/(5+q)-1/(4+q)}$,得到关于条件累积分布函数最优平滑参数的准确的最优速度。Li 和 Racine 所做的模拟结果表明这种方法表现很好。

184

6.2 平滑因变量情况下估计具有连续协变量的条件累积分布函数

在这一节我们讨论另一个也可以平滑因变量 Y 的估计量,正如我们在 1.4 节中提出的非条件累积分布函数的估计量那样。也就是说,我们可以估计 $F(y|x)$ 为

$$\hat{F}(y \mid x) = \frac{n^{-1} \sum_{i=1}^{n} G\left(\frac{y - Y_i}{h_0}\right) W_h(X_i, x)}{\hat{\mu}(x)} \qquad (6.3)$$

其中 $G(\cdot)$ 是由研究者选出的核累积分布函数,即标准正态累积分布函数,且 h_0 是与 Y 有关的平滑参数。

除了假定 6.1,我们也需要假定当 $n \to \infty$ 时,$h_0 \to 0$。定义 $\hat{M}(y,x) = [\hat{F}(y|x) - F(y|x)]\hat{\mu}(x)$,那么 $\hat{F}(y|x) - F(y|x) = \hat{M}(y,x)/\hat{\mu}(x)$。我们有以下结论。

定理 6.2 定义 $B_0(y,x) = (1/2)\kappa_2 F_{00}(y|x)$,并令 $\Omega(y,x) = \kappa^q F_0(y|x)/\mu(x)$。

注意 $\Sigma_{y|x}$ 和 $B_s(x)$ 与我们在定理 6.1 中的定义是一样的。令 $|\bar{h}|^2 = h_0^2 + \sum_{s=1}^{q} h_s^2 = \sum_{s=0}^{q} h_s^2$，那么，在给出和定理 6.1 相似的条件下，可得

(i) $E[\hat{M}(y,x)] = \mu(x)\sum_{s=0}^{q} h_s^2 B_s(y,x) + o(|\bar{h}|^2)$。

(ii) $\mathrm{var}[\hat{M}(y,x)] = (nH_q)^{-1}\mu(x)^2[\Sigma_{y|x} - h_0 C_k \Omega(y,x)] + o((nH_q)^{-1})$，

其中 $H_q = h_1 \cdots h_q$，$C_k = 2\int G(v)w(v)v\mathrm{d}v$。

(iii) 如果 $(nh_1\cdots h_q)^{1/2}\sum_{s=0}^{q} h_s^3 = o(1)$，那么

$$(nh_1\cdots h_q)^{1/2}\left[\tilde{F}(y|x) - F(y|x) - \sum_{s=1}^{q} h_s^2 B_s(y,x)\right] \xrightarrow{d} N\left(0, \Sigma_{y|x}\right)$$

定理 6.2 的证明在 6.9 节给出。

定理 6.2 意味着

$$\mathrm{MSE}[\hat{F}(y|x)] = \left[\sum_{s=1}^{q} h_s^2 B_s(y,x)\right]^2 + \frac{\Sigma_{y|x} - h_0 C_k \Omega(y,x)}{nh_1\cdots h_q} + (\mathrm{s.o.}) \tag{6.4}$$

185

我们可以选择平滑参数来最小化 $F(y|x)$ 的一个加权积分均方误差(WIMSE)的首项，WIMSE 由下式给定：

$$\mathrm{WIMSE} = \int\left\{\left[\sum_{s=0}^{q} h_s^2 B_s(y,x)\right]^2 + \frac{[\Sigma_{y|x} - h_0 C_k \Omega(y,x)]}{nh_1\cdots h_q}\right\} s(y,x)\mathrm{d}y\mathrm{d}x \tag{6.5}$$

其中 $\int \mathrm{d}x\mathrm{d}y = \sum_{x^d \in D}\int \mathrm{d}x^c \mathrm{d}y$。

如果人们使用窗宽选择的插入法，那么要首先估计 $B_s(y,x)$、$\Sigma_{y|x}$ 和 $\Omega(y,x)$，这要求人们选择初始的“试验”平滑参数。然而，计算一个 $q+1$ 维的积分的精确数值是非常困难的。但是，确实存在开发得很好的自动化数据驱动法，能够选择出平滑参数来估计条件累积分布函数。在缺乏理想的估计条件累积分布函数的数据驱动法的情形下，我们再次建议使用已经提出的用于条件累积分布函数估计的交错鉴定法。正如在第 5 章关于条件概率分布函数所讨论的，最优平滑给出 $h_0 \sim h_s \sim n^{-1/(5+q)}$（所有 $s=1,\cdots,q$）。下面我们证明对条件累积分布函数最优平滑要求 $h_0 = o(h_s)$ $(s=1,\cdots,q)$，以及 $h_s \sim n^{-1/(4+q)}$。为了清楚说明，我们首先考虑 $q=1$ 的情形，然后，再转向一般情形。

考虑 $q=1$ 的情形。由式(6.4)可知加权均方误差为

$$\begin{aligned}\text{WIMSE} &= \int \text{MSE}[\hat{F}(y \mid x)]s(y,x)\mathrm{d}y\mathrm{d}x \\ &= A_0h_0^4 + A_1h_0^2h_1^2 + A_2h_1^4 \\ &\quad + A_3(nh_1)^{-1} - A_4h_0(nh_1)^{-1} + O(\eta_n)\end{aligned} \tag{6.6}$$

186 其中 $\eta_n = h_1^6 + h_0^6 + (h_1^2 + h_0^2)(nh_1)^{-1}$，且 A_j 是一些常数，分别为：

$$A_0 = \int B_0(y,x)^2 s(y,x)\mathrm{d}y\mathrm{d}x$$

$$A_1 = 2\int B_0(y,x)B_1(y,x)s(y,x)\mathrm{d}y\mathrm{d}x$$

$$A_2 = \int B_1(y,x)^2 s(y,x)\mathrm{d}y\mathrm{d}x$$

$$A_3 = \Sigma_{y|x}s(y,x)\mathrm{d}y\mathrm{d}x$$

$$A_4 = \Omega(y,x)s(y,x)\mathrm{d}y\mathrm{d}x$$

所有的 A_j 除了 A_1 均为正，但是 A_1 可以为正、负或0。

一阶条件为

$$\frac{\partial\text{WIMSE}}{\partial h_0} = 4A_0h_0^3 + 2A_1h_0h_1^2 - A_4(nh_1)^{-1} + (\text{s.o.}) = 0 \tag{6.7}$$

$$\frac{\partial\text{WIMSE}}{\partial h_1} = 2A_1h_0^2h_1 + 4A_2h_1^3 - A_3(nh_1^2)^{-1} + (\text{s.o.}) = 0 \tag{6.8}$$

其中(s.o.)表示更低阶项。

从式(6.7)和式(6.8)中，我们能容易看出 h_0 和 h_1 不能有同样的阶数。并且，可以证明 h_0 必定有比 h_1 更低的阶。假定 $h_0 = c_0n^{-\alpha}$ 且 $h_1 = c_1n^{-\beta}$，那么我们必须有 $\alpha > \beta$。然后，由式(6.8)我们得到 $\beta = 1/5$，且将它代入式(6.7)得到 $\alpha = 2/5$。于是，最优平滑要求 $h_0 \sim n^{-2/5}$ 和 $h_1 \sim n^{-1/5}$。[①]

对于 $q > 1$ 的一般情形，由于对称性，所有的 h_s 应该具有相同的阶数，即 $h_s \sim n^{-\beta}(s = 1,\cdots,q)$ 以及 $h_0 \sim n^{-\alpha}$。那么容易证明 $\beta = 1/(4+q)$ 和 $\alpha = 2/(4+q)$。于是，最优的 $h_s \sim n^{-1/(4+q)}(s = 1,\cdots,q)$，且 $h_0 \sim n^{-2/(4+q)}$。

到目前为止，我们集中关注 $F(y|x)$ 的一个局部常数估计量。我们也可以用局部线性法估计 $F(y|x)$。正如我们先前讨论的，$F(y|x) = E[\mathbf{1}(Y_i \leqslant y) | X_i = x]$ 是在 $X_i = x$ 条件下 $\mathbf{1}(Y_i \leqslant y)$ 的条件均值函数。于是，我们也可以使用局部线性方法来估计这个条件均值函数，即通过 $\mathbf{1}(Y_i \leqslant y)$ 对 $(1,(X_i = x)')$ 回归并使用核权重。最终的截距估计量将会是 $F(y|x)$ 的局部线性估计量。可以证明，这个 187 估计量的偏误首项为

---

① 人们也可以尝试使用 $\alpha = \beta$，根据式(6.8)这将导致 $\beta = 1/5$，根据式(6.7)这将导致 $\beta = 1/4$，存在不一致。类似地，假定 $\alpha < \beta$ 也导致不一致。因此，我们必须有 $\alpha > \beta$。

$$(\kappa_2/2)\sum_{s=1}^{q} h_s^2 F_{ss}(y \mid x) \tag{6.9}$$

而方差首项为 $\Sigma_{y|x}$，这与定理 6.1 给出的局部常数估计量的结果是一样的。然而，正如 Hall, Wolff 和 Yao(1999)指出的，这个局部线性估计量具有两个不良的性质：(i) 它可能对 Y 不是单调的，(ii) 它可能不是被局限在[0,1]内取值。

受经验似然法的启发，Hall 等(1999)和 Cai(2002)引入一个加权局部常数估计量去克服因使用局部线性法而产生的问题。一个加权局部常数估计量为

$$\bar{F}(y \mid x) = \frac{\sum_{i=1}^{n} p_i(x) K_h(X_i, x) \mathbf{1}(Y_i \leqslant y)}{\sum_{i=1}^{n} p_i(x) K_h(X_i, x)} \tag{6.10}$$

其中 $p_i(x)(i=1,\cdots,n)$ 表示数据 $X_1,\cdots,X_n$ 的加权函数且设计点 x 满足性质 $p_i(x) \geqslant 0$，$\sum_{i=1}^{n} p_i(x) = 1$，以及

$$\sum_{i=1}^{n} (X_i - x) p_i(x) K_h(X_i, x) = 0 \tag{6.11}$$

条件(6.11)是由局部线性估计量产生的，这确保了估计偏差在边界和内部区域有相同的阶数。这也可以在 Taylor 扩展式中减小偏误首项的数值。然而，满足这些条件的 $p_i(x)$ 不是唯一确定的(因为 $p_i(x)$ 有 n 个参数，只有 $q+2$ 个约束条件)。Hall 等和 Cai 建议通过在这些约束条件下最大化 $\prod_{i=1}^{n} p_i(x)$ 选择 $p_i(x)$。这引出了下面的最优化问题

$$\max_{p_i(x)'s} \sum_{i=1}^{n} \ln p_i(x),$$

$$\text{s.t. 式(6.11)},\ p_i(x) \geqslant 0,\ \sum_{i=1}^{n} p_i(x) = 1 \tag{6.12}$$

令 γ 为与条件(6.11)相关的拉格朗日乘数，那么(见习题 6.5 的提示)

$$\hat{p}_i(x) = \frac{1}{n[1 + \hat{\gamma}(X_i - x) K_h(X_i, x)]} \tag{6.13}$$

其中 188

$$\hat{\gamma} = \arg\max_{\gamma} \frac{-1}{nh} \sum_{i=1}^{n} \ln[1 + \gamma(X_i - x) K_h(X_i, x)] \tag{6.14}$$

式(6.14)没有一个闭式解。Cai(2002)建议使用 Newton-Raphson 方法找到基于式(6.14)的 γ 的根(解)。在类似于定理 6.2 给出的那些正则性条件下，Cai 构造了 $\bar{F}(y|x)$ 的渐近分布为

$$(nh)^{1/2}\left[\bar{F}(y\mid x)-F(y\mid x)-\sum_{s=1}^{q}h_s^2\bar{B}_s(y,x)\right]\xrightarrow{d}N(0,\Sigma_{y\mid x})\quad(6.15)$$

其中 $\bar{F}(y|x)\bar{B}_s(y,x)=(1/2)\kappa_2F_{ss}(y|x)$。

式(6.15)证明 $\bar{F}(y|x)$ 与 $F(y|x)$ 的局部线性核估计量具有相同的(一阶)渐近分布。它还具有额外的优势,它对 Y 是单调的且它只在[0,1]取值。一个缺点是它的计算需要数值最优化程序。

当 X 有一个有界的支撑时,Cai(2002)进一步证明在 X 的支撑的边界上 $\bar{F}(y|x)$ 与偏误 $\left(\sum_{s=1}^{q}h_s^2\right)$ 和方差有同样的阶数,即这个方法避免了在有界支撑情况下的边界效应。

我们观察到上面的加权局部常数估计量具有类似于局部线性估计量的性质。上面提到一个缺点是其计算要求非线性优化程序,而局部常数和局部线性估计量拥有闭式解且容易被计算出来。使用局部线性估计量产生的问题是局部加权函数可能取负值,因此 $F(y|x)$ 的最终估计量可能取单位区间[0,1]之外的值。Hansen(2004)提出了一个修正的 $F(y|x)$ 的局部线性估计量,在那里他将其中的负值权重(如果存在的话)用0代替。这种修正是可以在渐近意义上忽略不计的,但是它限制 $F(y|x)$ 的最终局部线性估计量取值在[0,1]区间,从而得出一个有效的条件累积分布函数估计量。Hansen 证明他的修正的局部线性估计量的渐近分布与式(6.15)得出的 $\bar{F}(y|x)$ 是一阶等价的。

条件累积分布函数的非参数估计在经济学中存在广泛的潜在应用。比如说,这些估计量可以用于识别和估计非可加和非可分函数,参见 Matzkin(2003)以及 Altonji 和 Matzkin(2005)。

189

到目前为止,我们关注的是条件累积分布函数的非参数估计。当协变量的维数很高时,维数诅咒可能阻碍准确的非参数估计。Hall 和 Yao(2005)提出使用降维法去逼近条件概率密度函数。具体地说,他们使用 $F(Y|X'\beta)$ 去逼近真实的条件累积分布函数 $F(Y|X)$,这里 X 是 q 维的且 β 是 $q\times 1$ 维的未知参数向量。他们的估计方法在思想上是与单指数模型法类似的(见第8章)。因为 $X'\beta$ 是一个标量(即一元变量),它只包含一维的非参数估计,因此避免了维数诅咒的批评。

6.3 条件分位数函数的非参数估计

分位数回归方法已经被研究者广泛地应用。对于一系列参数模型的全面分析,我们推荐读者参考 Koenker(2005)。这种方法之所以流行的一个原因是它们更加充分地描述了一个感兴趣的变量的条件累积分布函数的特征,比较而

言回归模型仅仅给出了条件均值。

一个累积分布函数 $F(\cdot)$ 的无条件第 α 个分位数被定义为

$$q_\alpha = \inf\{y:F(y) \geqslant \alpha\} = F^{-1}(\alpha) \tag{6.16}$$

其中 $\alpha \in (0,1)$。

例如，令 $Y \sim N(0,1)$ 且令 $F(\cdot)$ 为相关的标准正态累积分布函数。那么 $q_{0.5}=5$，由于 $F(0)=0.5$ 且 $F(0-\varepsilon)<0.5$ 对任意 $\varepsilon>0$ 成立。为了得出分位数如 $q_{0.95}$，我们利用逆累积分布函数。也就是说，因为 $q_{0.95}=F^{-1}(0.95)$ 且 $F(q_{0.95})=F(F^{-1}(0.95))=0.95$，我们得到 $q_{0.95}=1.624$，因为 $P[N(0,1)\leqslant 1.645]=F(1.645)=0.95$。

一般地，如果令 X 表示一个累积分布函数为 $F(\cdot)$ 的随机变量，那么找到 q_α 的方法包括应用 F 到 $q_\alpha=F^{-1}(\alpha)$ 以得到

$$F(q_\alpha) \equiv P[X \leqslant q_\alpha] = \alpha$$

即应用 F 于式(6.16)。

再次通过例子，假定 X 表示某一年的家庭收入。那么 $q_{0.25}$ 是这样一种收入水平，即所有家庭中，有 25% 的家庭收入是低于这个值的。 190

通常，我们感兴趣的是条件分位数而不是非条件分位数。条件第 α 个分位数定义为($\alpha \in (0,1)$)

$$q_\alpha(x) = \inf\{y:F(y \mid x) \geqslant \alpha\} = F^{-1}(\alpha \mid x) \tag{6.17}$$

例如，假定 X 表示年龄而 Y 表示个人收入。那么 $q_{0.25}(x=30)$ 所指的收入水平的含义是：25% 的 30 岁的人的收入水平是低于这个值的。

在实践中，我们可以通过转换估计出的条件累积分布函数率估计条件分位数函数 $q_\alpha(x)$。分别利用对一个平滑函数的估计和一个关于 Y 的示性函数，我们可得

$$\hat{q}_\alpha(x) = \inf\{y:\hat{F}(y \mid x) \geqslant \alpha\} \equiv \hat{F}^{-1}(\alpha \mid x) \tag{6.18}$$

$$\hat{q}_\alpha(x) = \inf\{y:\tilde{F}(y \mid x) \geqslant \alpha\} \equiv \tilde{F}^{-1}(\alpha \mid x) \tag{6.19}$$

因为 $\hat{F}(y|x)$($\tilde{F}(y|x)$)位于 0 和 1 之间且对于 Y 是单调的，$\hat{q}_\alpha(x)$($\tilde{q}_\alpha(x)$)总是存在的。因此，一旦人们得到 $\hat{F}(y|x)$($\tilde{F}(y|x)$)，利用式(6.18)和式(6.19)计算 $\hat{q}_\alpha(x)$($\tilde{q}_\alpha(x)$)就是轻而易举的。

我们解释式(6.18)的意义是，对于 α 和 x 一个给定的值，我们通过选择 q_α 最小化以下目标函数解出 $q_\alpha(x)$

$$\hat{q}_\alpha(x) = \arg\min_q \mid \alpha - \hat{F}(q \mid x) \mid \tag{6.20}$$

也就是说，最小化式(6.20)的 q 的结果是 $\hat{q}_\alpha(x)$。

我们假定 $F(y|x)$ 有一个条件概率密度函数 $f(y|x)$，$f(y|x)$ 在 x^c 上是连续

的,且$f(q_\alpha(x)|x)>0$。注意因为$F_0(y|x)=\partial F(y|x)/\partial y$,有$f(y|x)\equiv F_0(y|x)$。

下面的两个定理给出了$\tilde{q}_\alpha(x)$和$\hat{q}_\alpha(x)$的渐近分布。

定理 6.3 对于$s=1,\cdots,q$,定义

$$B_{\alpha,s}(y,x)=\frac{B_s(y,x)}{f(q_\alpha(x)\mid x)}$$

其中$B_s(y,x)$是在定理6.1中被定义的。然后,与在定理6.1中给出的类似条件下,我们有

$$(nh_1\cdots h_q)^{1/2}\left[\tilde{q}_\alpha(x)-q_\alpha(x)-\sum_{s=1}^{q}h_s^2B_{\alpha,s}(x)\right]\xrightarrow{d}N(0,V_\alpha(x))$$

191 其中$V_\alpha(x)=\alpha(1-\alpha)\kappa^q/[f^2(q_\alpha(x)|x)\mu(q_\alpha(x))]$。

定理 6.4 定义$B_{\alpha,0}(x)=F_0(y|x)/\mu(x)$且$B_{\alpha,s}(y,x),s=1,\cdots,q$与定理6.3是一样的。与在定理6.2中给出的类似条件下,我们有

$$(nh_1\cdots h_q)^{1/2}\left[\hat{q}_\alpha(x)-q_\alpha(x)-\sum_{s=0}^{q}h_s^2B_{\alpha,s}(x)\right]\xrightarrow{d}N(0,V_\alpha(x))$$

其中$B_{\alpha,s}(s=1,\cdots,q)$且$V_\alpha(x)$与定理6.3中的定义是一致的。

定理6.3的证明留作习题,定理6.4的证明在6.9.1节给出。

也可以通过转换式(6.10)中给出的$F(y|x)$的加权局部常数估计量来估计$q_\alpha(x)$,即

$$\bar{q}_\alpha(x)=\inf\{y:\bar{F}(y\mid x)\geqslant\alpha\}\equiv\bar{F}^{-1}(\alpha\mid x)\tag{6.21}$$

Cai(2002)证明

$$(nh_1\cdots h_q)^{1/2}\left[\bar{q}_\alpha(x)-q_\alpha(x)-\sum_{s=0}^{q}h_s^2\bar{B}_s(q_\alpha(x),x)/f(x)\right]\xrightarrow{d}N(0,V_\alpha(x))\tag{6.22}$$

其中$\bar{B}_s(\cdot,\cdot)$是在式(6.15)中定义的且$V_\alpha(x)$与在定理6.4中的定义相同。

6.4 检验函数方法

$q_\alpha(x)$的另一个流行的估计量可以使用所谓的检验函数(check function)推导出来,参见Chaudhuri(1991),Chaudhuri,Doksum和Samarov(1997),Jones和Hall(1990),Yu和Jones(1997,1998),Honda(2000),Cheng和Peng(2002),Whang(2006),以及其中的参考文献。检验函数的命名来自潜在目标函数的形状。注意L_2规范目标函数,如基于最小二乘法的那些,得出的是U形目标函数,而L_1规范目标函数,如那些基于最小绝对差的那些,得出的是V形或者"检

192 验形"目标函数。

流行的参数线性分位数回归模型(Koenker 和 Bassett(1978))为以下形式

$$Y_i = X_i'\beta + u_i \tag{6.23}$$

从式(6.23)中可得 $q_\alpha(Y_i|X_i) = X_i'\beta + q_\alpha(u_i|X_i)$。我们不能分开识别 β 和 $q_\alpha(u_i|X_i)$,正如不能在最小二乘回归模型中分开识别截距和 $E(u_i)$ 那样(即此处我们强加 $E(u_i)=0$),我们可以将这个模型重写为

$$Y_i = q_\alpha(Y_i \mid X_i) + v_{\alpha,i} = X_i'\beta_\alpha + v_{\alpha,i} \tag{6.24}$$

其中 $v_{\alpha,i} = u_i - q_\alpha(u_i|X_i)$。定义 $q_\alpha(v_{\alpha,i}|X_i)=0$。基本上,我们强加了误差过程的条件第 α 个分位数等于 0 这个条件。

众所周知,人们可以通过最小化下面的目标函数估计式(6.24)中的 β_α:

$$\hat{\beta}_\alpha = \arg\min_\beta \sum_{i=1}^n \rho_\alpha(Y_i - X_i\beta) \tag{6.25}$$

其中 $\rho_\alpha(z) = z[\alpha - 1(z\leqslant 0)]$ 是检验函数,参见 Koenker 和 Bassett(1978)。同样,参见 He 和 Zhu(2003)对参数分位数回归模型的缺乏拟合检验(lack-of-fit test)。可以证明式(6.25)等价于

$$\hat{\beta} = \arg\min_\beta\left\{\alpha\sum_{Y_i\geqslant X_i'\beta}|Y_i - X_i'\beta| + (1-\alpha)\sum_{Y_i<X_i'\beta}|Y_i - X_i'\beta|\right\} \tag{6.26}$$

其中我们使用了当 $z<0$ 时,有 $-z=|z|$ 的结论。

Koenker 和 Bassett(1978)证明

$$\sqrt{n}(\hat{\beta}_\alpha - \beta_\alpha) \xrightarrow{d} N(0, \sigma_\alpha^2[E(X_iX_i')]^{-1})$$

其中 $\sigma_\alpha^2 = \alpha(1-\alpha)f_{v\alpha}(0)$,且 $f_{v\alpha}(\cdot)$ 是 v_α 的概率密度函数。

这个检验函数方法也可以用于非参数分位数回归模型的估计。我们使用一个局部常数模型(或者,我们可以选择使用一个局部线性模型),且我们选择 a 以最小化下面的目标函数:　193

$$\min_a \sum_{j=1}^n \rho_\alpha(Y_j - a)W_h(X_i, x) \tag{6.27}$$

其中 $\rho_\alpha(v) = v[\alpha - \mathbf{1}(v\leqslant 0)]$。令 $\hat{q}_{\alpha,lc}(x)$ 表示最小化式(6.27)的 a 值,可以证明 $\hat{q}_{\alpha,lc}(x)$ 是 $q_\alpha(x)$ 的一个一致估计量。

$\hat{q}_{\alpha,lc}(x)$ 的均方误差首项可以在 Jones 和 Hall(1990)以及 Yu 和 Jones(1997)中找到。$\hat{q}_{\alpha,lc}(x)$ 的偏误首项和方差首项与在定理 6.3 中给出的 $\tilde{q}_\alpha(x)$ 是完全相同的。因此,$\hat{q}_{\alpha,lc}(x)$ 与定理 6.3 中给出的 $\tilde{q}_\alpha(x)$ 具有相同的渐近分布。

也可以使用一个局部线性模型通过用下面的目标函数来替换式(6.27):

$$\min_{a,b} \sum_{j=1}^n \rho_\alpha(Y_j - a - (X_j - x)'b)W_h(X_j, x) \tag{6.28}$$

其中最小化目标函数的 a 是 $q_\alpha(x)$ 的局部线性估计量，且 b 估计了 $q_\alpha(x)$ 的导数。

令 $\hat{q}_{\alpha,ll}(x)$ 表示 $q_\alpha(x)$ 的最终估计量。$\hat{q}_{\alpha,ll}(x)$ 的 MSE 首项可以在 Fan, Hu 和 Truong(1994)以及 Yu 和 Jones(1997)中找到。方差首项与局部常数估计量的是一样的(因此与定理 6.3 给出的相同)，而偏误首项为

$$\text{bias}(\hat{q}_{\alpha,ll}(x)) = \frac{1}{2}\kappa_2 \sum_{s=1}^{q} h_s^2 q_{\alpha,ss}(x) \tag{6.29}$$

其中 $q_{\alpha,ss}(x) = \partial^2 q_\alpha(x)/\partial x_s^2$。

6.5 离散和连续协变量混合的条件累积分布函数和分位数估计

当 X 为混合离散和连续协变量的向量，我们需要用一个适合混合数据型的广义乘积核取代核函数。我们首先讨论在没有平滑相关变量 Y 的情况下估计条件累积分布函数。我们估计 $F(y|x)$ 为

194

$$\tilde{F}(y \mid x) = \frac{n^{-1}\sum_{i=1}^{n} \mathbf{1}(Y_i \leqslant y) K_\gamma(X_i, x)}{\hat{\mu}(x)} \tag{6.30}$$

其中 $\hat{\mu}(x) = n^{-1}K_\gamma(X_i, x)$ 是 $\mu(x)$ 的核估计量，

$$K_\gamma(X_i, x) = W_h(X_i^c, x^c) L_\lambda(X_i^d, x^d)$$

$$W_h(X_i^c, x^c) = \prod_{s=1}^{q} h_s^{-1} w((X_{is} - x_s)/h_s)$$

$$L_\lambda(X_i^d, x^d) = \prod_{s=1}^{r} l(X_{is}^d, x_s^d, \lambda_s)$$

$$l(X_{is}^d, x_s^d, \lambda_s) = \mathbf{1}(X_{is}^d = x_s^d) + \lambda_s \mathbf{1}(X_{is}^d \neq x_s^d)$$

如果我们选择平滑的因变量 Y，我们可得另一个 $F(y|x)$ 的估计量为

$$\hat{F}(y \mid x) = \frac{n^{-1}\sum_{i=1}^{n} G\left(\frac{y - Y_i}{h_0}\right) K_\gamma(X_i, x)}{\hat{\mu}(x)} \tag{6.31}$$

其中 $G(\cdot)$ 是一个累积分布函数，如标准正态累积分布函数，且 h_0 是与 Y 相关的平滑参数。

定义 $\tilde{M}(y,x) = [\tilde{F}(y|x) - F(y|x)]\hat{\mu}(x)$ 并令 $\hat{M}(y,x) = [\hat{F}(y|x) - F(y|x)]\hat{\mu}(x)$。下面的两个定理分别给出了 $\tilde{F}(y|x)$ 和 $\hat{F}(y|x)$ 的渐近分布。

定理 6.5 在与定理 6.1 给出的相似的正则性条件下，但是关于 x^c 的可微

性条件改变了(参见 Li 和 Racine(2008)),可得

(i) 令 $\eta_n = |\bar{h}|^6 + (nh_1\cdots h_q)^{-1}|\bar{h}|^2$,那么

$$\mathrm{MSE}[\widetilde{M}(y,x)] = \left\{ \sum_{s=1}^{q} h_s^2 B_{1s}(y \mid x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y \mid x) \right\}^2 + V_0(y \mid x)(nh_1\cdots h_q)^{-1} + O(\eta_n)$$

其中

$$B_{1s}(y,x) = (1/2)\kappa_2[2F_s(y \mid x)\mu_s(x) + \mu(x)F_{ss}(y \mid x)]$$

$$B_{2s}(y,x) = \mu(x)^{-1}\sum_{z^d \in \mathcal{S}^d} \mathbf{1}_s(z^d,x^d)[F(y \mid x^c,z^d)\mu(x^c,z^d) - F(y \mid x)\mu(x)]$$

其中 $\mathbf{1}_s(v^d,x^d)$ 在式(4.24)中定义。

(ii) 令 $\Sigma_{y|x} = F(y \mid x)[1 - F(y \mid x)]/\mu(x)$,那么

195

$$(nh_1\cdots h_q)^{1/2}\left[\tilde{F}(y \mid x) - F(y \mid x) - \sum_{s=1}^{q} h_s^2 B_{1s}(y \mid x) - \sum_{s=1}^{r} \lambda_s B_{2s}(y \mid x)\right] \xrightarrow{d} N(0,\Sigma_{y|x})$$

定理 6.6　定义 $|\lambda| = \sum_{s=1}^{r} \lambda_s$, $B_{10}(y \mid x) = (1/2)\kappa_2 F_{00}(y \mid x)$, $\Omega(y,x) = \kappa^q F_0(y \mid x)/\mu(x)$,且令 $\Sigma_{y|x}$, $B_{1s}(y,x)$ 和 $B_{2s}(y,x)$ 与定理6.5中的定义是一样的。那么,在 Li 和 Racine(2008)给出的假定下,可得

$$\mathrm{MSE}[\widetilde{M}(y,x)] = \mu(x)\left\{ \sum_{s=0}^{q} h_s^2 B_{1s}(y \mid x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y \mid x) \right\}^2 + \frac{\mu(x)^2[\Sigma_{y|x} - h_0\Omega(y,x)]}{nh_1\cdots h_q} + O(\eta_n)$$

其中 $\eta_n = |\bar{h}|^6 + |\lambda|^2 + (nh_1\cdots h_q)^{-1}(|\bar{h}|^2 + |\lambda|)$ 且

$$(nh_1\cdots h_q)^{1/2}\left[\hat{F}(y \mid x) - F(y \mid x) - \sum_{s=1}^{q} h_s^2 B_{1s}(y,x) - \sum_{s=1}^{r} \lambda_s B_{2s}(y \mid x)\right] \xrightarrow{d} N\left(0,\Sigma_{y|x}\right)$$

定理6.5和定理6.6的证明在6.9.2节给出。

如上所述,为了得到一个分位数估计量,转换 $\hat{F}(\cdot)$:

$$\hat{q}_\alpha(x) = \hat{F}^{-1}(\alpha \mid x)$$

在实践中我们通过下式计算 $\hat{q}_\alpha(x)$

$$\hat{q}_\alpha(x) = \arg\min_q |\alpha - \hat{F}(q \mid x)|$$

$\hat{q}_\alpha(x)$的渐近分布在下一个定理中给出。

定理 6.7　定义 $B_{n,\alpha}(x) = B_n(q_\alpha(x) \mid x)/f(q_\alpha(x) \mid x)$,其中

$(y = q_\alpha(x))B_n(y \mid x) = \left[\sum_{s=0}^{q} h_s^2 B_{1s}(y \mid x) + \sum_{s=1}^{r} \lambda_s B_{2s}(y \mid x)\right]$ 是 $\hat{F}(y|x)$ 的偏误首项。那么在定理 6.6 使用的相同的条件下,可得

196

(i) $\hat{q}_\alpha(x) \to q_\alpha(x)$ 依概率收敛。

(ii) $(nh_1 \cdots h_q)^{1/2}[\hat{q}_\alpha(x) - q_\alpha(x) - B_{n,\alpha}(x)]N(0, V_\alpha(x))$ 依分布收敛,其中 $V_\alpha(x) = \alpha(1-\alpha)\kappa^q/[f^2(q_\alpha(x)|x)\mu(q_\alpha(x))] \equiv V(q_\alpha(x)|x)/f^2(q_\alpha(x)|x)$(因为 $\alpha = F(q_\alpha(x)|x)$)。

定理 6.7 的证明与定理 6.4 的证明是类似的,因此在此省略。

在习题 6.4 中,读者被要求证明估计 $\hat{F}(y|x)$ 的最优平滑参数应该满足 $h_s \sim n^{-1/(4+q)}(s=1,\cdots,q)$, $\lambda_s \sim n^{-2/(4+q)}(s=1,\cdots,q)$,以及 $h_0 \sim n^{-2/(4+q)}$。遗憾的是,由于我们知识所限,当估计条件累积分布函数时,在加权积分均方误差最小的意义上,不存在一个自发的数据驱动方法可以选择最优的窗宽。给定条件累积分布函数和条件概率密度函数之间的密切关系,当选择窗宽且利用它们估计条件累积分布函数 $\hat{F}(y|x)$ 和条件分位数函数 $\hat{q}_\alpha(x)$ 时,Li 和 Racine (2008)推荐使用基于条件概率密度函数估计的最小二乘交错鉴定法。

令 $\bar{h}_s$ 和 $\bar{\lambda}_s$ 表示通过最小化概率分布函数交错鉴定函数选择的 h_s 和 λ_s 值(见第 5 章)。回忆在第 5 章我们证明最优窗宽阶数是 $\bar{h}_s \sim n^{-1/(5+q)}(s=0,1,\cdots,q)$;$\bar{\lambda}_s \sim n^{-2/(5+q)}(s=1,\cdots,r)$(假定 Y 为一个连续变量)。为了得到对 $\hat{F}(y|x)$ 有正确的最优速度的窗宽,Li 和 Racine(2008)建议使用 $\hat{h}_0 = \bar{h}_0 n^{\frac{1}{5+q}-\frac{2}{4+q}}$, $\hat{h}_s = \bar{h}_s n^{\frac{1}{5+q}-\frac{1}{4+q}}(s=1,\cdots,q)$ 和 $\hat{\lambda}_s = \bar{\lambda}_s n^{\frac{2}{5+q}-\frac{2}{4+q}}(s=1,\cdots,r)$。

对于上面的窗宽 $\hat{h}_0$、$\hat{h}_s$、$\hat{\lambda}_s$,定理 6.6 和定理 6.7 成立。

6.6 一个小的蒙特卡洛模拟研究

依据 Li 和 Racine(2008),我们简要地研究在式(6.18)和式(6.19)中定义的条件分位数估计量的有限样本表现。我们考虑的 DGP 为

$$Y_i = \beta_0 + \beta_1 X_{i1}^d + \beta_2 X_{i2}^d + \beta_3 \sin(X_i^c) + \varepsilon_i$$

其中 X_i^c 服从 $[-2\pi, 2\pi]$ 上的均匀分布,X_{i1}^d 和 X_{i2}^d 都是二项式(离散单元的总量是 $c=4$),且 $\varepsilon_i \sim N(0, \sigma^2)$,其中 $\sigma = 0.50$, $n=100$。

197

我们考虑在式(6.18)中给出的估计量,它使用一个示性函数处理 Y,即 $\mathbf{1}(y - Y_i \geq 0)$,并考虑在式(6.19)中给出的估计量,它使用平滑函数 $G((y-Y_i)/h_0)$,其中 $G(\cdot)$ 是高斯核。不管为 $G((y-Y_i)/h_0)$ 选择的核函数怎样,当 $h_0 \to 0$ 时,$G((y-Y_i)/h_0) \to \mathbf{1}(y - Y_i \geq 0)$,因此为了得出 $\mathbf{1}(y - Y_i \geq 0)$ 的

结果我们简单地设置 $h_y=0$（选择一个足够小的 h_0）。

我们做了1 000次蒙特卡洛试验，每次计算当 $\alpha=0.50$ 时的分位数估计量的均方误差。我们报告这1 000次试验中的相关均方误差的中位数，并研究基于一系列窗宽选择量的非平滑和平滑分位数法的相应表现。结果概括到在表6.1中，其中小于1的数值表明了平滑分位数估计量(6.19)的较优表现。

表6.1 非平滑和平滑分位数估计量的均方误差表现

h_{x^c}	$(\lambda_{x_1^d},\lambda_{x_2^d})$	$h_y=h_0$	MSE
$1.06\sigma_x n^{-1/6}$	0	示性	0.49
$1.06\sigma_x n^{-1/6}$	0	$1.06\sigma_y n^{-1/6}$	0.41
$1.06\sigma_x n^{-1/6}$	LSCV	$1.06\sigma_y n^{-1/6}$	0.33
LSCV	LSCV	LSCV	0.25

从表6.1中我们观察到，首先，通过用一个特别公式 $h_y=1.06\sigma_y n^{-1/6}$ 来平滑 Y，与使用一个 Y 的非平滑的示性函数得到的估计量相比，均方误差减小了20%。其次，通过最小二乘交错鉴定法平滑离散协变量使均方误差减小了20%。最后，通过最小二乘交错鉴定法（不是使用一个特别公式）选择 h_x 和 h_y 也使得均方误差进一步减小了25%。与 h_y 的特别选择相比，或者与因变量和离散协变量的示性函数相比，使用最小二乘交错鉴定法的均方误差减小了一半。注意在上面的试验中，所有的协变量都是相关的。在实践中，我们经常碰到的情况是一些协变量实际上是不相关的（即独立于因变量），在这种情况下，使用最小二乘交错鉴定法得到的效率甚至会更大，因为该方法可以自发地平滑掉非相关协变量。

具有相关数据的非参数分位数回归在Cai(2002)中有讨论。也可参见Koenker和Xiao(2002)，他们讨论了关于分位数回归过程的推断。 198

6.7 风险函数的非参数估计

把上面讨论的条件累积分布函数估计的结论和第5章讨论的条件概率密度函数估计量结合，可以容易地得到风险函数的非参数估计量。我们首先给出一个风险函数的正式定义。

定义6.1 令 T 表示一个随机变量，代表从某种状态退出的时间（如不再失业）。一个人在时间 t 之前属于一个状态，他在 t 后的 $\mathrm{d}t$ 这样一个短的时间长度区间内离开原状态的概率是 $P(t\leqslant T\leqslant t+\mathrm{d}t\mid T\geqslant t)$。风险函数(hazard function)被定义为

$$h(t) = \lim_{dt \to 0} \frac{P(t \leqslant T \leqslant t + dt \mid T \geqslant t)}{dt}$$

即瞬时退出率。

粗略地讲,风险函数可以理解为 $h(t)dt$ 是在 t 后的一个短的时间长度区间 dt 内从某种状态离开的概率,条件是在 t 时仍然是该状态。

风险函数经常用于描述在个人在现在(到目前这周为止)是失业的条件下,在一个紧随其后的将来时间段(如下周)退出失业状态的概率(即获得一份工作)。它们也经常用于描述一个病人在这周存活的条件下,在下个时期(如下周)死亡的概率。

可以证明风险函数可以被写为(见习题 6.6 的提示)

$$h(t) = \frac{f_T(t)}{1 - F_T(t)} = -\frac{d}{dt}\ln[1 - F_T(t)], \quad t > 0 \tag{6.32}$$

因此 $f_T(t) = h(t)\exp\left(-\int_0^t h(v)dv\right)$。

$1 - F_T(t)$ 称为"生存函数"。这是因为 $F_T(t) = P(T \leqslant t)$ 是在 t 期之前退出的概率,所以 $1 - F_T(t) = P(T > t)$ 是在 t 期之后生存的概率。

199 在风险和持久分析中人们普遍使用的参数概率密度函数是 Weibull 分布,其概率密度函数为

$$f_T(t; a, b) = abt^{b-1}\exp(-at^b), \quad t > 0,\ a, b > 0$$

风险函数为

$$h(x) = bat^{b-1}$$

因此风险函数在 $b > 1$ 下是严格递增的,在 $b < 1$ 下是严格递减的。当 $b = 1$ 时,Weibull 分布简化为具有常数风险的指数分布。Weibull 分布经常被用于对失业数据建模。

通常关注点在于在一组变量 x 的结果条件下的风险率。条件风险函数定义为

$$h(t \mid x) = \frac{f_T(t \mid x)}{1 - F_T(t \mid x)}, \quad t > 0 \tag{6.33}$$

比如说,T 可能是癌症患者死亡的时间,且 x 包含治疗的类型和个人的特征。我们假定 x 包含 q 个连续变量和 r 个离散变量。$h(t|x)$ 的非参数估计可以简单通过用它们的非参数估计量代替 $f_T(t|x)$ 和 $F_T(t|x)$ 而取得。

$$\hat{h}(t \mid x) = \frac{\hat{f}_T(t \mid x)}{1 - \hat{F}_T(t \mid x)} \tag{6.34}$$

其中 $\hat{f}_T(t|x)$ 和 $\hat{F}_T(t|x)$ 分别在第 5 章和 6.5 节中定义。由

$$\hat{F}_T(t \mid x) = F_T(t \mid x) + O_p\left(\sum_{s=1}^{q} h_s^2 + \sum_{s=1}^{q} \lambda_s + (nh_1 \cdots h_q)^{-1} \right)$$

容易看出 $\hat{h}(t|x)$ 的渐近分布是与 $\hat{f}_T(t|x)/a_{t,x}$ 相同的，其中 $a_{t,x} = 1 - F_T(t|x)$。因此，根据定理 6.3 我们马上得知

$$(nh_1 \cdots h_q)^{1/2} \left\{ \hat{h}(t \mid x) - h(t \mid x) - a_{t,x}^{-1} \left[\sum_{s=1}^{q} B_{1s}(t,x) h_s^2 + \sum_{s=1}^{r} B_{2,s}(t,x) \right] \right\}$$

$$\xrightarrow{d} N\left(0, \frac{V(t \mid x)}{a_{t,x}^2}\right) \tag{6.35}$$

其中 $B_{1s}(t,x)$ 和 $B_{2s}(t,x)$ 在定理 5.3 中定义。

200

在上面的讨论中我们假定数据不是截断的。在实际应用中，比如对受到特别治疗的病人的跟踪调查，一个病人可能在临终前没有对调查做出回应，数据将因此是截断的。我们将在第 11 章中讨论怎样在非参数和半参数背景中处理截断数据。

同样，当我们在高维背景中估计一个风险函数时，维数诅咒可能阻止了准确的非参数估计。在这种情形下，人们可以改为选择估计 Horowitz(1999)所建议的半参数风险模型，Horowitz 考虑了半参数比例风险模型。也可参考 Linton，Nielsen 和 van de Geer(2003)用基于核方法估计可乘和可加的风险模型的做法。

6.8　应　用

6.8.1　波士顿房屋数据

我们考察 20 世纪 70 年代的波士顿房屋数据，这一数据已被许多作者广泛分析过。接下来，我们报告在 Li 和 Racine(2007)中提出的应用。这个数据集包含 $n = 506$ 个观测数据，响应变量 Y 是在一个给定区域内房屋价格的中位数。根据 Chaudhuri 等(1997，p. 724)，我们关注三个重要的协变量：RM = 该区域每个房屋的平均房间数量(用最接近的整数表示)，LSTAT = 该区域低收入人口的百分比，DIS = 至波士顿五个就业中心的加权距离。一个有趣的特征是数据在 50 000 美元(20 年代房屋价格)处是右截断的，这使其十分适用于分位数方法。

我们首先打乱数据并创建两个独立的样本，样本容量分别为 $n_1 = 400$ 和 $n_2 = 106$。然后，我们使用容量为 n_1 的估计样本拟合一个线性参数分位数模型和非参数分位数模型，并基于容量为 n_2 的独立支持数据(hold-out data)中的协变量生成 Y 的预测中位数。最后我们计算均方预测误差(MSPE)，它被定义为

$$\text{MSPE} = n_2^{-1} \sum_{i=1}^{n_2} (Y_i - \hat{q}_{0.5}(X_i))^2$$

其中 $q_{0.5}(X_i)$ 是预测中位数，由参数模型或非参数模型生成，Y_i 是支持数据集对
201
应的真实值。为了防止潜在的批评，即我们的结果反映了一个不具代表性的划分数据，我们重复这个过程 100 次，每次计算相对 MSPE（即参数 MSPE 除以非参数 MSPE）。对于每次划分，我们采用 Hall 等（2004）的方法去计算数据相关窗宽。数据的 100 次划分的相对 MSPE 中位数是 1.13（下四分位数 = 1.03，上四分位数 = 1.20），表明非参数方法产生了优良的样本外分位数估计值。图 6.1 绘出了对于数据的 100 次随机划分，概括了这些结果的密度估计。

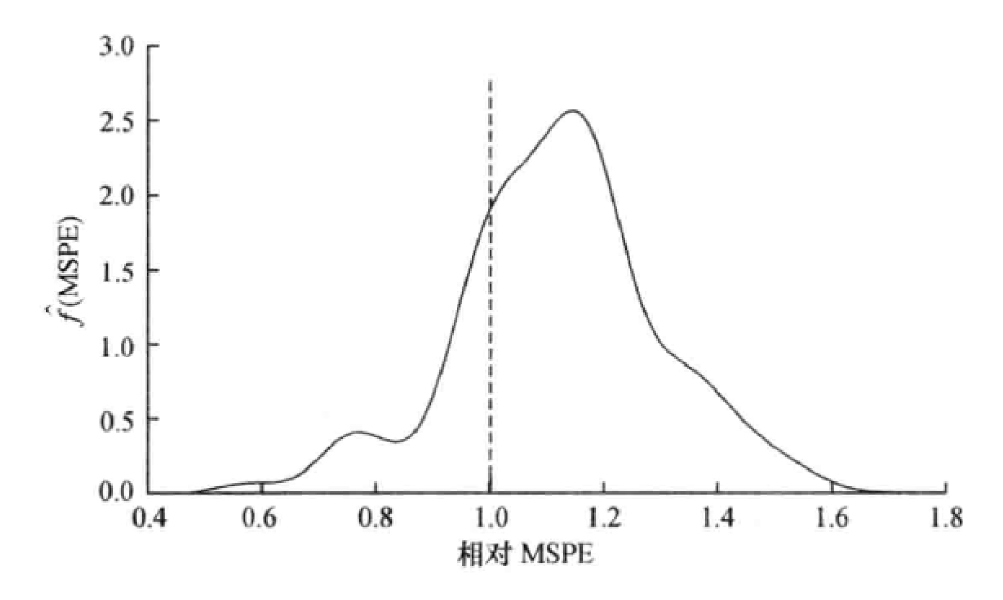


图 6.1　对数据的 100 次随机划分得到的相对 MSPE 的密度

注：值 > 1 表示对数据的一个给定随机划分，核方法有更好的样本外表现（下四分位数 = 1.03，上四分位数 = 1.20）。

图 6.1 显示出所有划分样本的 76%（即图中竖线的右边）相对效率大于或等于 1，在 100 次划分中有 76 次非参数分位数模型产生比参数分位数模型更好的房屋价格中位数预测。给定小样本容量和存在三个协变量的事实，我们感到这是一个有效的非参数方法的应用。当然，我们并不认为这会胜过一个近似正确的参数模型。更确切地说，我们仅仅希望我们可以比在文献中能够找到的普通参数设定做得更好。

202
6.8.2　青少年的生长曲线图

生长曲线图被儿科医师和父母们用于与一个标准范围比较孩子的生长。身高、体重、身体质量指数（BMI）测量值被用于记录按月年龄为基础的孩子的身高和体重。相同性别年龄孩子的测量值与标准或正常的范围作比较。

生长曲线图提供早期的预警，提示孩子有医学问题。例如，太快的生长可能表明脑积水的存在（头骨腔中有液体积聚），脑瘤或者其他原因会引起巨头畸形（有奇大的头，它与脑积水不同，因为没有颅内压力的增加），然而太慢的生长

可能表明存在大脑畸形、早期骨缝闭合或者其他问题。体重、身高或两者发育不足可能表明成长失调、慢性病、疏忽或其他问题。

以下应用的数据来自美国疾病控制预防中心的国家卫生预防和营养检验调查。各种方法被用于建立生长曲线,结果一般以分位数呈现出来。例如,美国疾病控制预防中心使用一个两阶段的方法。在第一阶段,经验百分位数被不同的参数和非参数方法平滑。为了得到相应的百分位数和 z 得分,平滑后的百分位数接着在第二阶段被修正后的最小中位数平方估计法近似估计。我们考察式(6.19)中所定义的估计量的混合数据分位数法的直接应用,通过例子构造年龄体重分位数。我们报告了第 25、50 和 75 个分位数并且在图 6.2 中绘制出男性的图。窗宽用最小二乘交错鉴定法得出,其中 $\hat{h}_{\text{weight}}=1.22$, $\hat{h}_{\text{age}}=8.11$, $\hat{\lambda}_{\text{sex}}=0.12$。

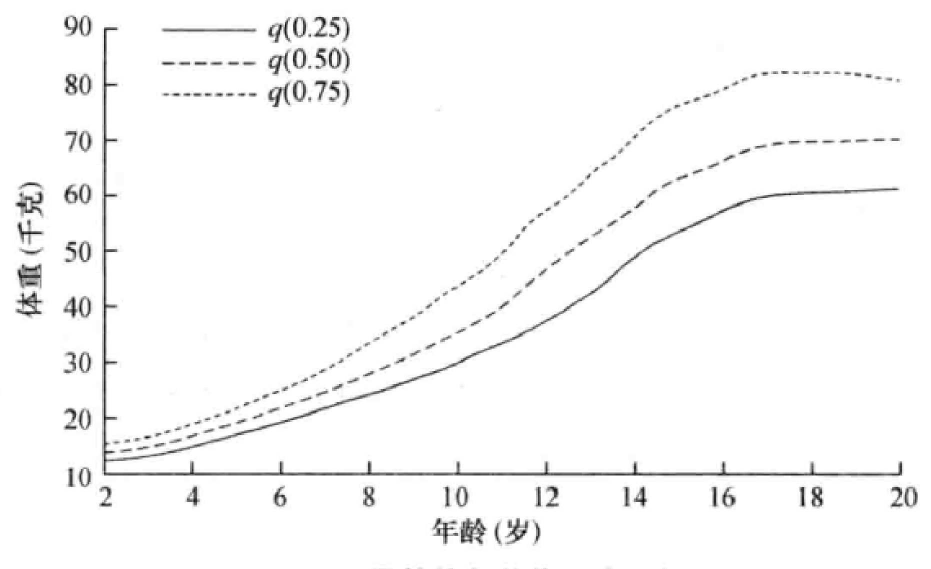


图 6.2　男性的年龄体重分位数

图 6.2 很好地揭示出数据集所呈现的许多特性,包括随年龄增长出现的异方差性和在体重的条件累积分布函数中的非对称性。

6.8.3　条件风险价值

金融工具受多种风险的影响,包括利息风险、违约风险、流动性风险和市场风险。风险价值(VaR)用来衡量后者。它是一个对以下损失数量的单独估计, 203 由于一个特定持有期的一般市场活动,在一种工具中人们的头寸可能在数量上会损失。人们可以把 VaR 当作对一给定的概率在一给定的时期内金融头寸的最大损失,因此,VaR 是在正常市场条件下由于一个意外事件带来的损失的一个简单指标(Tsay(2002, p. 257))。接下来,我们对 Li 和 Racine(2007)中讨论的应用进行介绍。

用 $\Delta V(l)$ 表示一种金融工具的价值从时间 t 到 $t+l$ 的变化，我们定义在 l 时间区间概率为 p 的 VaR 为

$$p = P[\Delta V(l) \leqslant \text{VaR}] = F_l(\text{VaR})$$

其中 $F_l(\text{VaR})$ 是 $\Delta V(l)$ 的未知累积分布函数。VaR 在 p 很小时取负值，因为金融工具持有者遭受损失。[①] 因此，持有者遭受的损失大于或者等于 VaR 的概率是 p，最大损失是 VaR 的概率是 $1-p$。$\text{VaR}_p = \inf\{\text{VaR} \mid F_l(\text{VaR}) \geqslant p\}$ 是 $F_l(\text{VaR})$ 的第 p 个分位数，因此 VaR_p 就是累积分布函数 $F_l(\text{VaR})$ 的第 p 个分位数。一般做法是去掉下标 p，例如“5% VAR”($\text{VaR}_{0.05}$)。显然累积分布函数 $F_l(\text{VaR})$ 在实践中是未知的，且必须被估计。

204

条件风险价值(CVaR)用来测量当含有协变量时的 VaR，它在一个协变量向量 X 的条件下进行估计。当然条件累积分布函数 $F_l(\text{VaR} \mid X)$ 同样未知并且必须被估计。在文献中能够找到各种各样的参数方法。接下来，我们把来自五个流行参数方法的结果与在式(6.19)中定义的非参数分位数估计量进行比较。

我们考虑在 Tsay(2002)中所用的 IBM 股票数据(r_t，从 1962 年 7 月 3 日到 1998 年 12 月 31 日日收益的对数(%))，如图 6.3 所示。收益的对数近似等于金融头寸价值的百分比变化，并且自始至终使用这种度量。CVaR 因此用百分比的形式，由 r_{t+1} 分布的分位数计算得出，以在 t 时可利用信息为条件，所以 CVaR 的金额等于个人头寸的现金价值乘以收益序列对数的 CVaR。接下来，我们计算一天区间内($l=1$)的 CVaR。

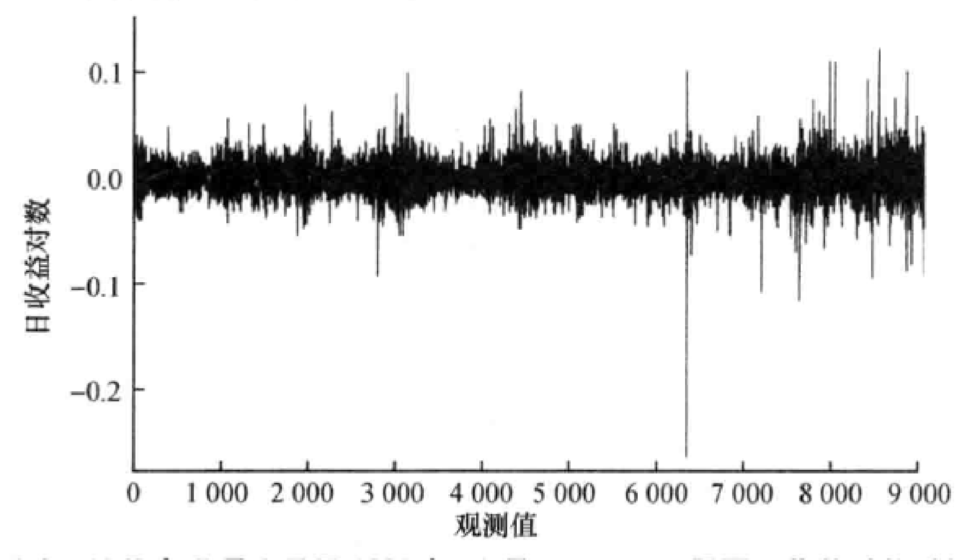


图 6.3　1962 年 7 月 3 日至 1998 年 12 月 31 日 IBM 股票日收益对数时间图

① 对一个长期头寸我们关心下尾回报，即由于价值下降的损失。对一个短期头寸我们关心上尾，即由于价值上升的损失。在后一种情况下我们简单地构建 $1-F_l(\text{VaR})$ 模型而不是 $F_l(\text{VaR})$。当然，在任何一种情况下我们都是在构建分布尾部的模型。

根据 Tsay(2002),我们使用下面的解释变量为 IBM 股票日收益对数的 CVaR 建模:

(i) X_{1t}:10 月、11 月、12 月示性变量,如果 t 在第四季度等于 1。这个变量关注第四季度的潜在影响(或者年末效应),如果存在,其对 IBM 日股票收益的影响。

(ii) X_{2t}:上一交易日行为示性变量,当且仅当上一交易日收益对数小于或等于 2.5% 时等于 1。这个变量表示在上一交易日当 IBM 股票价格下跌了 2.5% 或更多时恐慌出售的概率。

(iii) X_{3t}:波动率的定性测量,测量了在 $t-1$ 和 $t-5$ 之间对数收益数量超过一个阈值($|r_{t-i}| \geqslant 2.5\%$ $(i=1,2,\cdots,5)$)的元数。

(iv) X_{4t}:年度趋势,定义为(年限数 $t-1961$)/38,用来检测 IBM 股票极端收益的任何趋势行为。 205

(v) X_{5t}:对于均值修正序列,一个以高斯 GARCH(1,1)模型为基础的波动率序列,等于 σ_t,其中 σ_t^2 是 GARCH(1,1)模型的条件方差。

表 6.2 给出了 Tsay(2002,pp. 282,295)中报告的结果,并包含式(6.19)中定义的估计量,用许多现有的方法测算 1998 年 12 月 31 日的 CVaR。解释变量取值为 $X_{1,9190}=1$,$X_{2,9190}=0$,$X_{3,9190}=0$,$X_{4,9190}=0.9737$,$X_{5,9190}=1.7966$。这些值基于一个取值为 1 000 万美元的多头头寸,因此如果风险价值是 -2%,则我们得到 $\text{VaR}=10\,000\,000\times 0.02=200\,000$(美元)。

表 6.2　IBM 股票一个多头头寸的条件风险价值　　(单位:美元)

模型	5%	1%
异质泊松分布,GARCH(1,1)	303 756	497 425
条件正态分布,IGARCH(1,1)	302 500	426 500
AR(2)-GARCH(1,1)	287 700	409 738
学生 t_5 分布,AR(2)-GARCH(1,1)	283 520	475 943
极值	166 641	304 969
LSCV CDF	258 727	417 192

可以观察到,根据对参数模型的不同选择,在这个例子中对 5% CVaR 人们可以得到的估计值相差 82% 之多,对 1% CVaR 相差 63%。当然,这个差异由模型的不确定性产生。有人也许会用贝叶斯方法去处理这种不确定性,或者用平均模型,或者另考虑对函数设定稳健的非参数方法。核分位数方法提供了可能对实际工作者有价值的有意义估计的指导。非参数分位数估计量的另一个用 206

处也许是非参数地评估累积分布函数,在通过以上的参数估计量产生的值处来评价参数值事实上对应哪个分位数,当然假定参数模型是误设的。

有趣的是,Tsay(2002)发现 X_{1t} 和 X_{2t} 是不相关的解释变量。最小二乘交错鉴定法窗宽与之相一致。然而,X_{3t} 也被发现是不相关的,因此只有一个波动性指标是相关的。

6.8.4 意大利真实收入,1951—1998 年

我们再次考虑之前在 1.13.5 节中讨论过的意大利 GDP 面板数据,它包含意大利 21 个地区 1951—1998 年的收入数据(百万里拉,1990 年为基期)。我们考虑在式(6.19)中定义的估计量并对第 25、50 和 75 收入分位数建模,把时间看成有序离散回归元。图 6.4 绘出了分位数结果估计值。

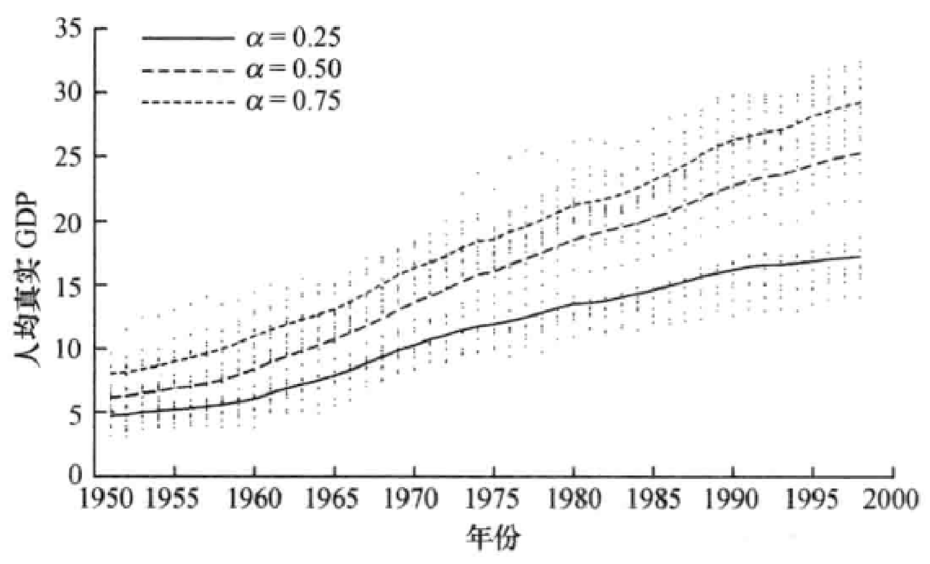


图 6.4 意大利真实 GDP 的收入分位数

回忆第 1 章图 1.9 表明收入分布从 20 世纪 50 年代早期的单峰曲线发展为 20 世纪 90 年代的显著双峰。这个特征被非参数分位数估计量清晰地捕捉,正如在图 6.4 中所看到的。

6.8.5 多元 Y 条件累积分布函数示例:OECD 条件下的 GDP 和人口增长

在第 5 章 5.5.5 节我们对多元 Y 条件概率密度函数建模,用窗宽选择的数据驱动方法,样本容量是 $n=616$,Y 由下面的变量组成:y_1,每阶段人均收入增长率;y_2,每阶段年人口增长率;x,OECD 身份(0/1)。我们现在估计条件累积分布函数 $F(y_1,y_2|x)$,并展示在图 5.5 中给出的条件概率密度函数的结果,即我们再次用 Hall 等(2004)最小平方交错鉴定法绘出 $\hat{F}(y_1,y_2|X=0)$ 和 $\hat{F}(y_1,y_2|X=1)$。

207

图 6.5 给出了一个多元 Y 条件累积分布函数的清晰画面。一个随机占优

关系是明显的,OECD 国家与非 OECD 国家相比倾向于拥有更低的人口增长率和更高的 GDP 增长率。

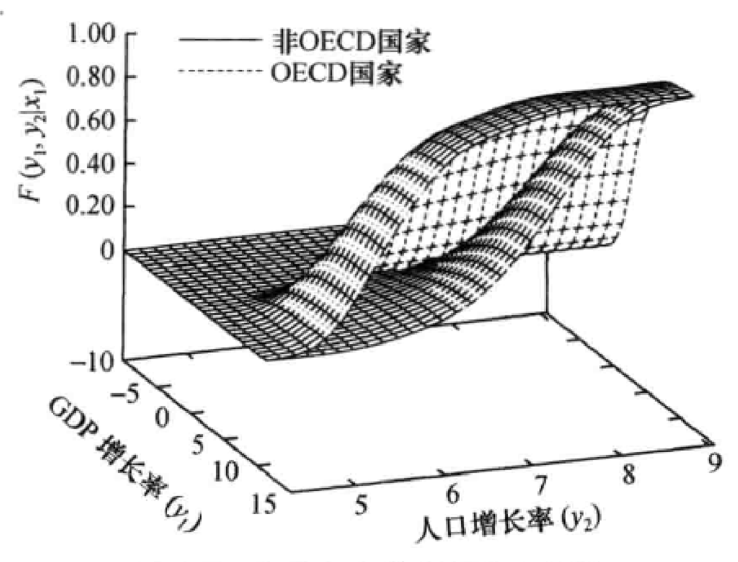


图 6.5 多元 Y 条件累积分布函数

6.9 证 明

209

6.9.1 定理 6.1、6.2 和 6.4 的证明

定理 6.1(i)的证明:

$$
\begin{aligned}
E[\tilde{F}(y\mid x)\hat{\mu}(x)] &= E[\mathbf{1}(Y_i \leqslant y)W_h(X_i,x)] \\
&= E[W_h(X_i,x)E(\mathbf{1}(Y_i \leqslant y)\mid X_i)] \\
&= E[W_h(X_i,x)F(y\mid X_i)] \\
&= (h_1\cdots h_q)^{-1}\int\mu(z)W\left(\frac{z-x}{h}\right)F(y\mid z)\,\mathrm{d}z \\
&= \int\mu(x+hv)F(y\mid x+hv)W(v)\,\mathrm{d}v \\
&= \int\left[\mu(x)+\sum_{s=1}^{q}\mu_s(x)h_sv_s+\frac{1}{2}\sum_{s=1}^{q}\sum_{t=1}^{q}\mu_{ts}(x)h_sh_tv_sv_t\right] \\
&\quad\times\left[F(y\mid x)+\sum_{s=1}^{q}F_s(y\mid x)h_sv_s\right. \\
&\quad\left.+\frac{1}{2}\sum_{s=1}^{q}\sum_{t=1}^{q}F_{ts}(y\mid x)h_sh_tv_sv_t\right]W(v)\,\mathrm{d}v+o(|h|^2) \\
&= \mu(x)F(y\mid x)+\frac{1}{2}\kappa_2\sum_{s=1}^{q}h_s^2[\mu(x)F_{ss}(y\mid x)
\end{aligned}
$$

$$+ 2\mu_s(x)F_s(y \mid x) + F(y \mid x)\mu_{ss}(x)] + o(|h|^2) \qquad (6.36)$$

类似地，

$$E[\hat{\mu}(x)] = E[W_h(X_i,x)]$$
$$= \mu(x) + (1/2)\kappa_2 \sum_{s=1}^{q} h_s^2 \mu_{ss}(x) + o(|h|^2) \qquad (6.37)$$

注意 $E[\tilde{M}(y,x)] = E[\tilde{F}(y|x)\hat{\mu}(x)] - F(y|x)E[\hat{\mu}(x)]$，所以定理 6.1(i) 可以从式(6.36)和式(6.37)中得到。 □

210 定理 6.1(ii)的证明：

$$\begin{aligned}
\mathrm{var}(\tilde{M}(y,x)) &= n^{-1}\mathrm{var}[(\mathbf{1}(Y_i \leq y) - F(y \mid x))W_h(X_i,x)] \\
&= n^{-1}\{E[(\mathbf{1}(Y_i \leq y) - F(y \mid x))^2]W_h(X_i,x)^2 + O(1)\} \\
&= n^{-1}E\{E[(\mathbf{1}(Y_i \leq y) - F(y \mid x))^2 \mid X_i]W_h(X_i,x)^2\} + O(n^{-1}) \\
&= n^{-1}E\{[F(y \mid X_i) - 2F(y \mid x)F(y \mid X_i) \\
&\quad + F(y \mid x)^2]W_h(X_i,x)^2\} + O(n^{-1}) \\
&= n^{-1}\int \mu(z)[F(y \mid z) - 2F(y \mid x)F(y \mid z) \\
&\quad + F(y \mid x)^2]W_h(z,x)^2 dz + O(n^{-1}) \\
&= (nh_1 \cdots h_q)^{-1}\int \mu(x - hv)[F(y \mid x - hv) - 2F(y \mid x)F(y \mid x - hv) \\
&\quad + F(y \mid x)^2]W(v)^2 dv + O(n^{-1}) \\
&= \kappa^q (nh_1 \cdots h_q)^{-1}\mu(x)[F(y \mid x) - F(y \mid x)^2] \\
&\quad + O((nh_1 \cdots h_q)^{-1} |h|^2 + n^{-1})
\end{aligned}$$

其中 $\kappa = \int w(v)^2 dv$。 □

定理 6.1(iii)的证明：定义 $B_n(y,x) = \sum_{s=1}^{q} h_s^2 B_s(y,x)$，那么

$$\begin{aligned}
&[\tilde{F}(y \mid x) - F(y \mid x) - B_n(y,x)] \\
&\quad = [\tilde{F}(y \mid x) - F(y \mid x) - B_n(y,x)]\hat{\mu}(x)/\hat{\mu}(x) \\
&\quad = [\tilde{F}(y \mid x) - F(y \mid x) - B_n(y,x)]\hat{\mu}(x)/\mu(x) + o_p(1) \\
&\quad \equiv \tilde{A}(y,x)/\mu(x) + o_p(1)
\end{aligned}$$

其中 $\tilde{A}(y,x) = [\tilde{F}(y|x) - F(y|x) - B_n(y,x)]\hat{\mu}(x)$。由定理 6.1(i)和(ii)我们可知：$E[\tilde{A}(y,x)] = O(|h|^3)$，以及 $\mathrm{var}[\tilde{A}(y,x)] = (nh_1 \cdots h_q)^{-1}\mu(x)^2\mathrm{var}(y,x) + O(\eta_n)$ $(\eta_n = (nh_1 \cdots h_q)^{-1}|h|^2)$。通过中心极限定理，可以证明：

$$(nh_1 \cdots h_q)^{1/2}\tilde{A}(y,x) \to N\left(0, \mu(x)^2 \Sigma_{y|x}\right) \text{ 依分布收敛}$$

因此，

$$(nh_1\cdots h_q)^{1/2}[\bar{F}(y,x)-F(y,x)-B_n(y,x)]$$
$$=(nh_1\cdots h_q)^{1/2}\bar{A}(y,x)/\mu(x)+o_p(1)$$
$$\xrightarrow{d}\mu(x)^{-1}N(0,\mu(x)^2\Sigma_{y|x})=N(0,\Sigma_{y|x})\qquad\square$$

定理 6.2(i)的证明：使用引理 6.1 我们有　211

$$
\begin{aligned}
E[\hat{M}(y,x)] &= E\left[G\left(\frac{y-Y_i}{h_0}\right)W_h(X_i,x)\right]\\
&= E\left\{E\left[G\left(\frac{y-Y_i}{h_0}\right)\mid X_i\right]W_h(X_i,x)\right\}\\
&= E\left\{\left[F(y\mid X_i)+\frac{\kappa_2h_0^2}{2}F_{00}(y\mid X_i)\right]W_h(X_i,x)\right\}+o(h_0^2)\\
&= \int\mu(z)\left[F(y\mid z)+\frac{\kappa_2h_0^2}{2}F_{00}(y\mid z)\right]W_h(z,x)\,\mathrm{d}z+o(h_0^2)\\
&= \int\mu(x+hv)\left[F(y\mid x+hv)\right.\\
&\quad\left.+\frac{\kappa_2h_0^2}{2}F_{00}(y\mid x+hv)\right]W(v)\,\mathrm{d}v+o(h_0^2)\\
&= \mu(x)F(y\mid x)+\frac{\kappa_2h_0^2}{2}\mu(x)F_{00}(y\mid x)+\kappa_2\sum_{s=1}^{q}h_s^2\mu_s(x)F_s(y\mid x)\\
&\quad+\frac{\kappa_2}{2}F(y\mid x)\sum_{s=1}^{q}\mu_{ss}(x)+o(|\bar{h}|^2)
\end{aligned}\tag{6.38}
$$

由式(6.37)和式(6.38)可以得到定理 6.2(i)。　$\square$

定理 6.2(ii)的证明：令 $H_q=h_1\cdots h_q$，通过使用引理 6.1，我们有　212

$$
\begin{aligned}
&\mathrm{var}[\hat{M}(y,x)-F(y\mid x)\hat{\mu}(x)]\\
&= n^{-1}\mathrm{var}\left[\left(G\left(\frac{y-Y_i}{h_0}\right)-F(y\mid x)\right)W_h(X_i,x)\right]\\
&= E\left\{E\left[G^2\left(\frac{y-Y_i}{h_0}\right)\middle|X_i\right]-2F(y\mid x)E\left[G\left(\frac{y-Y_i}{h_0}\right)\middle|X_i\right]\right.\\
&\quad\left.+F(y\mid x)^2W_h^2(X_i,x)\right\}+O(1)\\
&= E\{[F(y\mid X_i)-h_0C_kF_0(y\mid X_i)-2F(y\mid X_i)F(y\mid x)\\
&\quad+F(y\mid x)^2+O(|\bar{h}|^2)]W_h^2(X_i,x)\}+O(n^{-1})\\
&= \int\mu(z)[F(y\mid z)-h_0C_kF_0(y\mid z)+2F(y\mid z)F(y\mid x)
\end{aligned}
$$

$$+ F(y \mid x)^2] W_h^2(z,x) \mathrm{d}z + (\mathrm{s.o.})$$

$$= (nH_q)^{-1} \int \mu(x + hv) [F(y \mid x + hv) - h_0 C_k F_0(y \mid x + hv)$$

$$+ 2F(y \mid x + hv) F(y \mid x) + F(y \mid x)^2] W^2(v) \mathrm{d}v + O(n^{-1})$$

$$= (nH_q)^{-1} \mu(x) \kappa^q \{ F(y \mid x) [1 - F(y \mid x)] - h_0 C_k \mu(x) F_0(y \mid x) \}$$

$$+ O(\mid \bar{h} \mid^2 (nH_q)^{-1} + n^{-1})$$

$$\equiv (nH_q)^{-1} \mu(x)^2 \left[\Sigma_{y|x} - h_0 \Omega(y,x) \right] + O(\mid \bar{h} \mid^2 (nH_q) - 1 + n^{-1}) \quad \square$$

定理 6.2(ii)的证明：根据(i)、(ii)以及 Liapunov 中心极限定理，可得(iii)。 □

引理 6.1

(i) $E\left[G\left(\frac{y - Y_i}{h_0} \right) \middle| X_i \right] = F(y|X_i) + (1/2) \kappa_2 h_0^2 F_{00}(y|X_i) + o(h_0^2)$。

(ii) $E\left[G^2\left(\frac{y - Y_i}{h_0} \right) \middle| X_i \right] = F(y|X_i) - h_0 C_k F_0(y|X_i) + o(h_0)$，其中 $C_k = 2\int G(v) w(v) v \mathrm{d}v$。

证明：本题证明留作习题。 □

引理 6.2 $\hat{E}(y + \varepsilon_n | x) - \hat{F}(y | x) = f(y|x)_{\varepsilon_n} + O_p(h_0^2) + o_p(\varepsilon_n + (nh_1 \cdots h_q)^{-1/2})$。

213

证明：令 $A_n(\varepsilon_n) = [\hat{F}(y + \varepsilon_n | x) - \hat{F}(y|x)] \hat{\mu}(x) / \mu(x)$，那么 $\hat{F}(y + \varepsilon_n | x) - \hat{F}(y|x) = A_n(\varepsilon_n)[1 + o_p(1)]$。通过引理 6.1，我们有

$$E[A_n(\varepsilon_n)] = E\left\{ E\left[G\left(\frac{y + \varepsilon_n - Y_i}{h_0} \right) - G\left(\frac{y - Y_i}{h_0} \right) \middle| X_i \right] \frac{K_\gamma(X_i, x)}{\mu(x)} \right\}$$

$$= E\{ [F(y + \varepsilon_n \mid x) - F(y \mid x) + O(h_0^2)] K_\gamma(X_i, x) \}$$

$$= E\{ [f(y \mid X_i) \varepsilon_n + O(\varepsilon_n^2) + O(h_0^2)] K_\gamma(X_i, x) \}$$

$$= f(y \mid X_i) \varepsilon_n / \mu(x) + O(\varepsilon_n^2 + h_0^2)$$

类似地，

$$\mathrm{var}[A_n(\varepsilon_n)] \leqslant n^{-1} \mu(x)^{-2} E\left\{ \left[G\left(\frac{y + \varepsilon_n - Y_i}{h_0} \right) - G\left(\frac{y - Y_i}{h_0} \right) \right]^2 \times K_\gamma^2(X_i, x) \right\}$$

$$= O(\varepsilon_n (nh_1 \cdots h_q)^{-1})$$

因此，

$$\hat{F}(y + \varepsilon_n \mid x) - \hat{F}(y \mid x) = A_n(\varepsilon_n)[1 + o_p(1)]$$

$$= f(y \mid x)\varepsilon_n + O_p(h_0^2) + o_p(\varepsilon_n + (nh_1 \cdots h_q)^{-1/2}) \qquad \square$$

定理 6.4 的证明：由定理 6.2 我们知道 $\hat{F}(y|x) \to F(y|x)$ 依概率收敛。由 Tucker(1967) 的定理 1 得

$$\sup_{y \in \mathbb{R}} | \hat{F}(y \mid x) - F(y \mid x) | \to 0 \text{ 依概率收敛} \tag{6.39}$$

因为 $F(y|x)$ 是一个条件累积分布函数。

由于 $q_\alpha(x)$ 是唯一的，这意味着

$$\delta = \delta(\varepsilon) = \min\{\alpha - F(q_\alpha(x) - \varepsilon \mid x),\quad F(q_\alpha(x) + \varepsilon \mid x) - \alpha\} > 0$$

容易得到

$$\begin{aligned} P[| \hat{q}_\alpha(x) - q_\alpha(x) | > \varepsilon] &\leqslant P[| F(\hat{q}_\alpha(x) \mid x) - F(y \mid x) | > \delta] \\ &\leqslant P[\sup_y | \hat{F}(y \mid x) - F(y \mid x) | > \delta] \end{aligned} \tag{6.40}$$

由式(6.39)它趋于 0。因此，$\hat{F}(y|x) - F(y|x) \xrightarrow{p} 0$。我们现在来证明定理 6.4。 214
对任意 v，令 $\varepsilon_n = B_\alpha(x) + (nh)^{-1/2}\sigma_\alpha(x)v$，那么由引理 6.2 和假定 $h_0 = o((nh_1 \cdots h_q)^{1/2})$，

$$\begin{aligned} Q_\alpha(v) &\stackrel{\text{def}}{=} P[(nH^q)^{1/2}\sigma_\alpha^{-1}(x)[\hat{q}_\alpha(x) - q_\alpha(x) - B_{n,\alpha}(x) \\ &\qquad + o_p(|h|^2 + |\lambda|)] \leqslant v] \\ &\sim P[\hat{q}_\alpha(x) \leqslant q_\alpha(x) + \varepsilon_n] \\ &= P[\hat{F}(q_\alpha(x) + \varepsilon_n \mid x) \geqslant \alpha] \\ &\sim P[\hat{F}(q_\alpha(x) \mid x) \geqslant -f(q_\alpha(x) \mid x)\varepsilon_n + \alpha] \end{aligned} \tag{6.41}$$

因此，

$$\begin{aligned} Q_n(v) \sim & P[(nh_1 \cdots h_q)^{1/2}\sigma^{-1}(q_\alpha(x) \mid x) \\ & \times \{\hat{F}(q_\alpha(x) \mid x) - \alpha - B_n(q_\alpha(x) \mid x)\} \geqslant -v] \sim \Phi(v) \qquad \square \end{aligned}$$

其中 $\Phi(\cdot)$ 是标准正态分布。

6.9.2　定理 6.5 和 6.6（混合协变量情形）的证明

定理 6.5 由下面的引理 6.3 和 6.4 证明，而定理 6.6 由引理 6.5 和 6.6 证明。

引理 6.3

(i) $E[\widetilde{M}(y,x)] = \mu(x)\sum_{s=1}^{q} h_s^2 B_{1s}(y,x) + \mu(x)\sum_{s=1}^{r} \lambda_s B_{2s}(y,x) + o(|\lambda| + |h|^2)$。

(ii) $\text{var}(\widetilde{M}(y,x)) = \kappa^q \mu(x) F(y|x)[1 - F(y|x)]$，其中 $B_{1s}(y,x)$ 和 $B_{2s}(y,x)$

在定理6.5中定义。

证明:见 Li 和 Racine(2007)。

引理 6.4

$$(nh_1\cdots h_q)^{1/2}\left[\tilde{F}(y\mid x)-F(y\mid x)-\sum_{s=1}^{q}h_sB_{1s}(y,x)-\sum_{s=1}^{r}\lambda_sB_{2s}(y,x)\right]\to N(0,V)$$

其中 $V=\kappa^qF(y|x)[1-F(y|x)]/\mu(x)$。

215

证明:注意 $\tilde{F}(y|x)-F(y|x)=\tilde{M}(y,x)/\hat{\mu}(x)=\tilde{M}(y,x)/\mu(x)+o_p(1)$,引理6.4由引理6.3和 Liapunov 中心极限定理得到。

引理 6.5

(i) $E[\hat{M}(y,x)]=\mu(x)\sum_{s=0}^{q}h_s^2B_{1s}(y,x)+\mu(x)\sum_{s=1}^{r}\lambda_sB_{2s}(y,x)+o(|\lambda|+|h|^2)$。

(ii) $\mathrm{var}(\hat{M}(y,x))=\kappa^q(nh_1\cdots h_q)^{-1}\mu(x)\times[F(y|x)-F(y|x)^2\sim h_0C_kF_0(y|x)]$。

证明:这里的证明留作习题。 □

引理 6.6

$$(nh_1\cdots h_q)^{1/2}\left[\hat{F}(y\mid x)-F(y\mid x)-\sum_{s=1}^{q}h_s^2B_{1s}(y,x)-\sum_{s=1}^{r}\lambda_sB_{2s}(y,x)\right]\xrightarrow{d}N(0,V)$$

其中 $V=\kappa^q[F(y|x)-F(y|x)^2]/\mu(x)$。

证明:注意

$$\hat{F}(y\mid x)-F(y\mid x)=\hat{M}(y,x)/\hat{\mu}(x)=\hat{M}(y,x)/\mu(x)+o_p(1)$$

引理6.6由引理6.5和 Liapunov 中心极限定理得到。 □

6.10 习 题

习题 6.1 证明引理6.1。

提示:用变量替换和分部积分,如我们在1.4节的做法。也要注意

$$2\int G(v)w(v)\mathrm{d}v=\int \mathrm{d}G^2(v)=G^2(v)\mid_{-\infty}^{\infty}=1$$

习题 6.2 证明引理6.5。

习题 6.3 证明定理6.3。

提示:模仿定理6.4的证明,但不要引入 λ_0。

习题6.4 证明对一般的 q 和 r 情形,最小化 IMSE$[\hat{F}(y|x)]$ 的最优平滑参数要求 $h_s \sim n^{-1/(4+q)}(s=1,\cdots,q)$,$\lambda_s \sim n^{-2/(4+q)}(s=1,\cdots,r)$,$h_0 \sim n^{-2/(4+q)}$。 216

提示:使用的论证与式(6.6)和式(6.7)下面的两段相似。

习题6.5 推导式(6.13)。

提示:拉格朗日函数记为

$$\mathcal{L} = \sum_{i=1}^{n} \ln p_i(x) - \gamma_1 \left[\sum_{i=1}^{n} p_i(x) - 1 \right] - \gamma_2 \sum_{i=1}^{n} (X_i - x) p_i(x) K_h(X_i, x)$$

一阶条件有

$$\frac{\partial \mathcal{L}}{\partial \gamma_1} = \sum_{i=1}^{n} p_i(x) - 1 \overset{\text{set}}{=} 0 \tag{6.42}$$

$$\frac{\partial \mathcal{L}}{\partial \gamma_2} = \sum_{i=1}^{n} p_i(x)(X_i - x) K_h(X_i, x) \overset{\text{set}}{=} 0 \tag{6.43}$$

$$\frac{\partial \mathcal{L}}{\partial p_i(x)} = \frac{1}{p_i(x)} - \gamma_1 - \gamma_2 (X_i - x) K_h(X_i, x) \overset{\text{set}}{=} 0 \tag{6.44}$$

式(6.43)可以导出 $1 = p_i(x)[\gamma_1 + \gamma_2(X_i - x)K_h(X_i, x)]$。关于 i 求和,再由式(6.42)和式(6.43)得

$$n = \gamma_1 \sum_{i=1}^{n} p_i(x) + \gamma_2 \sum_{i=1}^{n} p_i(x)(X_i - x) K_h(X_i, x) = \gamma_1$$

将 $\gamma_1 = n$ 代入式(6.44),我们得到($\gamma_2 = \gamma$)

$$p_i(x) = \frac{1}{n[1 + \gamma(X_i - x)K_h(X_i, x)]}$$

这证明了式(6.13)。

因此,我们可以通过选择 γ 来最大化对数似然函数:

$$\mathcal{L}(\gamma) = \sum_{i=1}^{n} \ln p_i(x) = -\sum_{i=1}^{n} \ln[1 + \gamma(X_i - x)K_h(X_i, x)]$$

习题6.6 217

(i) 推导式(6.32)。

提示:

$$\begin{aligned}
h(t) &= \lim_{\Delta t \to 0} \left(\frac{1}{\Delta t} P(t \leqslant T < t + \Delta t \mid T \geqslant t) \right) \\
&= \lim_{\Delta t \to 0} \left(\frac{1}{\Delta t} \int_t^{t+\Delta t} f(v \mid T \geqslant t) \, dv \right) \\
&= \lim_{\Delta t \to 0} \left(\frac{1}{\Delta t} \int_t^{t+\Delta t} f_T(v) \, dv / P(T \geqslant t) \right) \\
&= f_T(t) / P(T \geqslant t) \\
&= f_T(t) / [1 - F_T(t)]
\end{aligned}$$

(ii) 证明 $f_X(x) = h(x)\exp\left(-\int_0^x h(y)\,\mathrm{d}y\right)$。

提示：

$$\begin{aligned}\int_0^y h(v)\,\mathrm{d}v &= \int_0^v \frac{f_T(v)}{1-F_T(v)}\mathrm{d}v \\ &= -\ln[1-F_T(v)]\mid_{v=0}^{v=t} \\ &= -\ln[1-F_T(v)]\end{aligned}$$

这意味着

$$1-F_T(t) = \mathrm{e}^{-\int_0^t h(v)\,\mathrm{d}v}$$

代入(i)，我们得到(ii)。

第 2 部分　半参数方法

第7章　半参数的部分线性模型

在这一章我们要讨论一个相对简单且已广泛使用的半参数模型——半参数部分线性回归模型。可以参考 Härdle, Liang 和 Gao(2000)对部分线性模型的一个全面处理。大致来说,半参数模型就是模型中某些成分为参数,而其余成分为未设定的函数。因此,这些模型包括有限维参数,同时也包括一些未知函数,这些未知函数可以被视为含有无限维参数的函数。 221

部分线性模型是实践中所使用的最简单的半参数模型之一。我们将用这个模型来介绍半参数模型,因为它的估计是直接的,只涉及回归函数的基本核估计和最小二乘回归。这个模型也用于说明在半参数模型的估计中所产生的一些微妙的问题。例如,有限维参数(模型的参数部分)通常用一个参数$\sqrt{n}$速度来估计,尽管我们通常需要更强的正则性条件和对平滑参数更加严格的条件以得到模型参数部分的$\sqrt{n}$速度。

7.1　部分线性模型 222

一个半参数部分线性模型由下式给出:

$$Y_i = X_i'\beta + g(Z_i) + u_i, \quad i = 1, \cdots, n \quad (7.1)$$

其中 X_i 是 $p \times 1$ 阶向量,β 是 $p \times 1$ 阶未知参数的向量,并且 $Z_i \in \mathbb{R}^q$。$g(\cdot)$的函数形式未设定。有限维参数 β 组成模型的参数部分,未知函数 $g(\cdot)$组成模型的非参数部分。数据被假定为独立同分布并且 $E(u_i | X_i, Z_i) = 0$,我们也允许存在一个未知形式的条件异方差误差过程 $E(u_i^2 | x, z) = \sigma^2(x, z)$。我们的讨论关注怎样得到 β 的

一个$\sqrt{n}$一致估计量,一旦得到,$g(\cdot)$的估计量也可以很容易地得到。

7.1.1 β 的识别

为了识别参数向量β,需要一些识别条件。注意 X 不能包含常数(即 β 不能包含截距),因为如果包含截距,例如 α,它不能从未知函数 $g(\cdot)$ 分离地识别出来。也就是说,对任意常数 $c\neq0$,观察到 $\alpha+g(z)=[\alpha+c]+[g(z)-c]\equiv\alpha_{new}+g_{new}(z)$,于是,新的截距与新的函数 $g(\cdot)$之和与式(7.1)中旧的和在观测上相等。因为 $g(\cdot)$的函数形式未设定,这立刻告诉我们在一个部分线性模型中,截距项不能被识别。我们在 7.2 节推导出 β 的半参数估计量的渐近分布后,可以明显看到,β 的识别条件变成要求 $\Phi\overset{\text{def}}{=}E\{[X-E(X|Z)][X-E(X|Z)]'\}$ 是一个正定矩阵,这意味着 X 不能包含一个常数并且 X 的组成部分不能是 Z 的一个确定性的函数,否则 $X-E(X|Z)\equiv0$ 且 Φ 将是奇异的。

7.2 罗宾逊的估计量

我们首先用式(7.1)的一个不可行估计量去描述 β 估计中所涉及的机理。以 Z_i 为条件对式(7.1)取期望,我们得到

$$E(Y_i\mid Z_i)=E(X_i\mid Z_i)'\beta+g(Z_i)\tag{7.2}$$

223

从式(7.1)减去式(7.2)得到:

$$Y_i-E(Y_i\mid Z_i)=(X_i-E(X_i\mid Z_i))'\beta+u_i\tag{7.3}$$

定义 $\tilde{Y}_i=Y_i-E(Y_i|Z_i)\tilde{X}_i=X_i-E(X_i|Z_i)$,并对式(7.3)采用最小二乘法,我们得到一个 β 的估计量

$$\hat{\beta}_{\inf}=\left[\sum_{i=1}^{n}\tilde{X}_i\tilde{X}_i'\right]^{-1}\sum_{i=1}^{n}\tilde{X}_i\tilde{Y}_i\tag{7.4}$$

通过 Lindeberg-Levy 中心极限定理我们立即可以得到(见习题7.1)

$$\sqrt{n}(\hat{\beta}_{\inf}-\beta)\xrightarrow{d}N(0,\Phi^{-1}\Psi\Phi^{-1})\tag{7.5}$$

给定 Φ 是正定的,其中 $\Psi=E[\sigma^2(X_i,Z_i)\tilde{X}_i\tilde{X}_i']$,$\Phi=E(\tilde{X}_i\tilde{X}_i')$。

这个过程的基本思想是首先用式(7.1)减去式(7.2)来消除未知函数 $g(\cdot)$。然而,虽然未知函数 $g(\cdot)$在式(7.3)中被消除了,但是又引入了两个新的未知函数,即 $E(Y_i|Z_i)$和 $E(X_i|Z_i)$。因此,上面的估计量 $\hat{\beta}_{\inf}$不可行,因为 $E(Y_i|Z_i)$和 $E(X_i|Z_i)$是未知的。然而,我们知道这些条件期望使用核方法可以被一致估计,所以我们用它们的核估计量去代替 $\hat{\beta}_{\inf}$中出现的未知条件期望,因此得到一个 β 的可行估计量。也就是说,我们分别用 $Y_i-\hat{Y}_i$ 和 $X_i-\hat{X}_i$ 去代

替$\tilde{Y}_i = Y_i - E(Y_i | Z_i)$和$\tilde{X}_i = X_i - E(X_i | Z_i)$，其中

$$\hat{Y}_i \equiv \hat{E}(Y_i \mid Z_i) \overset{\text{def}}{=} n^{-1} \sum_{j=1}^{n} Y_j K_h(Z_i, Z_j) / \hat{f}(Z_i) \tag{7.6}$$

$$\hat{X}_i \equiv \hat{E}(X_i \mid Z_i) \overset{\text{def}}{=} n^{-1} \sum_{j=1}^{n} X_j K_h(Z_i, Z_j) / \hat{f}(Z_i) \tag{7.7}$$

$$\hat{f}(Z_i) = n^{-1} \sum_{j=1}^{n} K_h(Z_i, Z_j) \tag{7.8}$$

其中$K_h(Z_i, Z_j) = \prod_{s=1}^{q} h_s^{-1} k\left(\frac{z_{is} - z_{js}}{h_s}\right)$。

在推导β可行估计量的渐近分布时，随机分母$\hat{f}(Z_i)$的存在会导致一些技术上的困难。我们将考虑两种不同的方法来解决随机分母的存在，其中一种方法是在随机分母较小时通过使用一个函数来“修剪”观测值，另一种是使用密度加权的方法完全消除随机分母。我们以修剪法的讨论开始，定义一个β的可行估计量为 224

$$\hat{\beta} = \left\{ \sum_i (X_i - \hat{X}_i)(X_i - \hat{X}_i)' \right\}^{-1} \sum_i (X_i - \hat{X}_i)(Y_i - \hat{Y}_i) \mathbf{1}_i \tag{7.9}$$

其中$\mathbf{1}_i = \mathbf{1}(\hat{f}(Z_i) \geqslant b)$，当$\hat{f}(Z_i) \geqslant b$时该值等于 1，否则等于 0；修剪参数$b = b_n > 0$且当$n \to \infty$时满足$b_n \to 0$。

为了推导$\hat{\beta}$的渐近分布，我们首先提供一种定义并做出一些假定。我们将使用$\mathcal{G}_\nu^\alpha$，其中$\alpha > 0$且$\nu \geqslant 2$是一个整数，来表示这样的平滑函数类：如果$g \in \mathcal{G}_\nu^\alpha$，则$g$是$\nu$次可微的；$g$和它的偏导函数（直到$\nu$阶）都满足 Lipschitz 型条件，如$|g(z) - g(z')| \leqslant H_g(z) \| z' - z \|$，其中$H_g(z)$是一个具有有限$\alpha$矩的连续函数，且$\| \cdot \|$代表欧几里得范数，即$\| z \| = \sqrt{\sum_{j=1}^{q} z_j^2}$。

条件 7.1

(i) (Y_i, X_i, Z_i)，$i = 1, 2, \cdots, n$是独立同分布观测值，Z_i服从概率密度函数$f \in \mathcal{G}_\nu^\alpha$（即$f$是有界的），$g \in \mathcal{G}_\nu^4$，$E(X|z) \in \mathcal{G}_\nu^4$，其中$\nu \geqslant 2$是一个整数。

(ii) $E(u|X, Z) = 0$，$E(u^2 | x, z) = \sigma^2(x, z)$是关于$z$连续的，$X$和$\nu$都有有限四阶矩。

(iii) $K(\cdot)$是一个乘积核，单变量核$k(\cdot)$是一个有界ν阶核，且$k(v) = O(1/[1 + |v|]^{\nu+1})$。

(iv) 当$n \to \infty$时，$n(h_1 \cdots h_q)^2 b^4 \to \infty$，$nb^{-4} \sum_{s=1}^{q} h_s^{4\nu} \to 0$。

条件 7.1(i)陈述了一系列平滑和矩条件。未知函数$g(z)$和$E(X|z)$被假

225 定为 ν 阶可微。这些加上条件 7.1(iii) 的 ν 阶核，确保了核估计量的偏误是 $O\left(\sum_{s=1}^{q} h_s^{\nu}\right)$ 阶的。

Robinson(1988)曾用到条件 7.1(iv)。在实证应用中，人们可以忽略修剪参数 b，因为可以让 b 以极慢的速度趋于零。这样，条件 7.1(iv)基本等价于

$$\sqrt{n}\left[\sum_{s=1}^{q} h_s^{2\nu} + (nh_1 \cdots h_q)^{-1}\right] \to 0, \quad \text{当 } n \to \infty \text{ 时}$$

该条件很容易理解：$O\left(\sum_{s=1}^{q} h_s^{2\nu} + (nh_1 \cdots h_q)^{-1}\right)$ 是非参数均方误差的阶。可行估计量 $\hat{\beta}$ 和不可行估计量 $\hat{\beta}_{\inf}$ 之差与非参数估计误差平方的均值成比例。因此，由于 $\hat{\beta}$ 是 β 的 $\sqrt{n}$ 一致估计量，人们需要估计误差项的平方比 $n^{-1/2}$ 小，即 $\sum_{s=1}^{q} h_s^{2\nu} + (nh_1 \cdots h_q)^{-1}$ 是 $o(n^{-1/2})$ 阶的，导致了条件 7.1(iv)。

定理 7.1　在条件 7.1 下，我们有

$$\sqrt{n}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \Phi^{-1}\Psi\Phi^{-1}) \tag{7.10}$$

假定 Φ 是正定的，其中

$$\Phi = E[\tilde{X}_i \tilde{X}_i'], \quad \Psi = E[\sigma^2(X_i, Z_i)\tilde{X}_i \tilde{X}_i'], \quad \tilde{X}_i = X_i - E(X_i \mid Z_i)$$

定理 7.1 的证明可在 Robinson(1988)中找到。$\hat{\beta}$ 的渐近方差的一致估计量在习题 7.1 中给出。

比较定理 7.1 与式(7.5)给出的分布结果，我们看出可行估计量 $\hat{\beta}$ 与不可行估计量 $\hat{\beta}_{\inf}$ 有完全相同的渐近分布。这种结果后面的直觉很简单。如果我们忽视了修剪参数 b，那么，本质上 $\hat{\beta} - \hat{\beta}_{\inf} = O_p\left(\sum_{s=1}^{q} h_s^{2\nu} + (nh_1 \cdots h_q)^{-1}\right)$。根据条件 7.1(iv)，非参数估计量的均方误差速度是 $o_p(n^{-1/2})$。因此 $\sqrt{n}(\hat{\beta} - \hat{\beta}_{\inf}) = o_p(1)$，这意味着这两个估计量有相同的渐近分布。

假定人们使用一个二阶核，即假定 $\nu = 2$。由于 $h_s \sim h$①，条件 7.1(iv)变为 $n^{1/2}\left[\sum_{s=1}^{q} h_s^4 + (nh_1 \cdots h_q)^{-1}\right] \sim n^{1/2}[h^4 + (nh^q)^{-1}] = o(1)$，这要求 $q < 4$(或 $q \leqslant$

226 3，因为 q 表示 Z 的维数，是一个正整数)。这就是 Robinson(1988)使用的条件。

因此，如果 $q \geqslant 4$，条件 7.1(iv)要求人们使用一个更高阶的核。然而，Li(1996)证明 7.1(iv)可用一个下面所给出的更弱的条件替代。

---

① $h_s \sim h$ 意味着 $(h_s - h)/h = o(1)$，或等价地，$h_s = h + o(h)$。

条件 7.2　当 $n\to\infty$ 时，我们有 $nb^{-4}(h_1\cdots h_q)^2\Big/\left(\sum_{s=1}^{q}h_s^4\right)\to\infty$，$nb^{-4}(h_1\cdots h_q)\to\infty$，且 $nb^4\sum_{s=1}^{q}h_s^{4\nu}\to 0$。

Li(1996)证明

$$\hat{\beta}-\hat{\beta}_{\text{inf}}=O_p\left(\sum_{s=1}^{q}h_s^{2\nu}+\sum_{s=1}^{q}h_s^2(nh_1\cdots h_q)^{-1}+n^{-1/2}(nh_1\cdots h_q)^{-1}\right)$$

如果此项是 $o_p(n^{-1/2})$ 阶，我们可得到条件 7.2。关于估计误差为什么有这个阶而不是更为熟悉的 $O_p\left(\sum_{s=1}^{q}h_s^{2\nu}+(nh_1\cdots h_q)^{-1}\right)$ 阶，一个详细的解释可参见 Li(1996)（也可参见习题 7.5 中进一步的解释）。

考虑所有的 h_s 有相同阶数的情形（对所有的 s 有 $h_s\sim h$），则条件 7.2 变成 $n\max\{h^{2q-4},h^q\}\to\infty$ 且 $nh^{4\nu}\to 0$。如果人们使用一个二阶核（$\nu=2$），这个条件要求 $4\nu=8>\max(2q-4,q)$，进而要求 $q<6$，或 $q\leqslant 5$，因为 q 为一个正整数。因此，只要 $q\leqslant 5$，一个非负二阶核能够产生 β 的$\sqrt{n}$一致估计量。

下面的推论的证明可参见 Li(1996)。

推论 7.1　除条件 7.1(iv)由条件 7.2 代替外，在定理 7.1 给定的相同条件下，定理 7.1 仍然成立。

上文所述的半参数估计量的一个不令人满意的特点是用到修剪函数，这要求研究者选择一个冗余的参数，修剪参数 b。然而，人们可用如下的密度加权方法替代，来避免核估计量中的随机分母。

用 $f_i=f(Z_i)$ 乘以式(7.3)，我们可得到

$$(Y_i-E(Y_i\mid Z_i))f_i=(X_i-E(X_i\mid Z_i))'\beta f_i+u_if_i \tag{7.11}$$

可用最小二乘法估计 β，通过 $(Y_i-E(Y_i|Z_i))f_i$ 对 $(X_i-E(X_i|Z_i))f_i$ 回归。令 $\hat{\beta}_{\text{inf},f}$ 表示 β 的不可行估计量，则根据 Lindeberg-Levy 中心极限定理（见习题 7.1），我们知道 $\hat{\beta}_{\text{inf},f}$ 是$\sqrt{n}$一致且渐近正态的，即

227

$$\sqrt{n}(\hat{\beta}_{\text{inf},f}-\beta)\xrightarrow{d}N(0,\Phi_f^{-1}\Psi_f\Phi_f^{-1}) \tag{7.12}$$

其中，$\Phi_f=E[\tilde{X}_i\tilde{X}_i'f_i^2]$ 且 $\Psi_f=E[\sigma^2(X_i,Z_i)\tilde{X}_i\tilde{X}_i'f_i^4]$，其中 $\tilde{X}_i=X_i-E(X_i|Z_i)$。

β 的一个可行估计量可通过用 $(Y_i-\hat{Y}_i)f_i$ 和 $(X_i-\hat{X}_i)f_i$ 替代 $(Y_i-E(Y_i\mid Z_i))f_i$ 和 $(X_i-E(X_i\mid Z_i))f_i$ 得到，其中 $\hat{Y}_i$、$\hat{X}_i$ 和 $\hat{f}_i$ 是式(7.6)到式(7.8)中所定义的 $E(Y_i\mid Z_i)$、$E(X_i\mid Z_i)$ 和 $f_i(X_i)$ 的核估计量。因为 $\hat{Y}_i\hat{f}_i=n^{-1}\sum_i Y_iK_h(Z_i,Z_j)$（和 $\hat{X}\hat{f}_i$）不含有随机分母($\hat{f}_i$)，在消去修剪参数 b 后，条件 7.1(iv) 可由下面的条件代替。

条件 7.3 当 $n \to \infty$ 时，$n(h_1 \cdots h_q)^2 \to \infty$，且 $n\sum_{s=1}^{q} h_s^{4\nu} \to \infty$，则条件 7.2 可被替换为

条件 7.4 当 $n\to\infty$ 时，

$$n\min\left\{ (h_1\cdots h_q) \Big/ \left(\sum_{s=1}^{q} h_s^4 \right), (h_1\cdots h_q)^2 \right\} \to \infty$$

且 $n\sum_{s=1}^{q} h_s^{4\nu} \to 0$。

下一个定理给出了可行密度加权估计量的渐近分布。

定理 7.2 令 $\hat{\beta}_f$ 表示 $\hat{\beta}$ 的可行密度加权估计量，那么除条件 7.1(iv) 由条件 7.3 或 7.4 代替外，在定理 7.1 给出的同样的条件下，我们有

$$\sqrt{n}(\hat{\beta}_f - \hat{\beta}) \xrightarrow{d} N(0, \Phi_f^{-1}\Psi_f\Phi_f^{-1})$$

其中 Φ_f 和 Ψ_f 在式(7.12)中定义。

Li(1996)在较弱的条件 7.4 下证明了定理 7.2。在 7.5 节中我们使用更强的条件 7.3 提供了一个证明。正如将看到的，更强的条件 7.3 可产生一个相对简单的证明。

定理 7.2 表明可行估计量 $\hat{\beta}_f$ 与不可行估计量 $\hat{\beta}_{\inf,f}$ 有相同的渐近分布。

228

注意在估计 β 时，为避免随机分母而使用特定加权函数 $f(Z_i)$ 并未基于任何有效的论证。事实上，当误差项存在条件同方差时，未加权估计量 $\hat{\beta}$ 可以被证明是半参数有效的。

当误差项存在条件异方差时，即 $E(u_i^2|X_i,Z_i)=\sigma^2(X_i,Z_i)$，人们可能会有这样类似的思考，通过选择 $w_i=1/\sigma(X_i,Z_i)$ 作为权函数来得到 β 的一个有效估计量。然而，一般而言，这种直觉是不正确的。这种方法将不会产生 β 的有效估计，除了在特殊的情况下，条件方差仅是 Z_i 的一个函数。也就是说，仅当 $E(u_i^2|X_i,Z_i)=\sigma^2(Z_i)$ 时，选择 $1/\sigma(Z_i)$ 为权重将导致 β 的有效估计。在一般情况下，β 的有效估计更加复杂，并将在 7.4 节讨论。

7.2.1 非参数函数的估计

从式(7.2)中我们知道 $g(Z_i)=E(Y_i-X_i'\beta|Z_i)$。因此，在获得 β 的一个 $\sqrt{n}$ 一致估计量(即 $\hat{\beta}$)之后，$g(z)$ 的一致估计量可表示为

$$\hat{g}(z) = \frac{\sum_{j=1}^{n}(Y_j - X_j'\hat{\beta})K_h(z,Z_j)}{\sum_{j=1}^{n}K_h(z,Z_j)} \tag{7.13}$$

其中 $\hat{\beta}$ 可被替换为 $\hat{\beta}_f$。我们知道非参数核估计量的收敛速度比参数$\sqrt{n}$速度慢。因此，很容易看出，$\hat{g}(z)$渐近地等于如下利用到 β 真值的不可行估计量：

$$\tilde{g}(z) = \frac{\sum_{j=1}^{n}(Y_j - X_j'\beta)K_h(z,Z_j)}{\sum_{j=1}^{n}K_h(z,Z_j)} \tag{7.14}$$

$\tilde{g}(z)$的收敛速度和渐近分布在第2章讨论，从 $\tilde{g}(z)$的收敛速度和渐近分布可以立即得到 $\hat{g}(z)$的渐近分布（见习题7.2）。

注意估计 $g(z)$时对 h_s 的选择与估计 β 时的情形很不一样。为了得到 β 的$\sqrt{n}$一致估计量，如果 $q\geqslant6$，一个更高阶的核是必需的。然而，当估计 $g(z)$时，不论 q 取何值，都没有必要用到更高阶的核。因此，人们总可以用一个非负二阶核来估计 $g(z)$，且可以通过最小二乘交错鉴定法（在估计 $g(z)$时）来选择平滑参数，即人们总是可以选择 $h_1,\cdots,h_q$ 来最小化

$$\sum_{i=1}^{n}[Y_i - X_i'\hat{\beta} - \hat{g}_{-i}(Z_i,h)]^2 \tag{7.15}$$

其中 $\hat{g}_{-i}(Z_i,h)=\hat{g}_{-i}(Z_i)$如式（7.13）所定义的，其中 z 被 Z_i 替代，且 $\sum_{j=1}^{n}$ 被 $\sum_{j=1,j\neq i}^{n}$ 替代。

注意在式（7.15）中的因变量是 $Y_i - X_i'\hat{\beta}$ 而不是 Y_i。因为 $\hat{\beta}-\beta=O_p(n^{-1/2})$ 比任何非参数估计量有更快的收敛速度，可以把式（7.15）中的 $\hat{\beta}$ 替换为 β 来研究交错鉴定法选择的 h_s 的渐近行为。h_s 的收敛速度与第2章讨论的是一样的。

可以通过最小化以下式子来同时选择 $\hat{\beta}$ 和 h。

$$\sum_{i=1}^{n}[Y_i - X_i'\beta - \hat{g}_{-i}(Z_i,h)]^2$$

在一般条件下，包括使用一个二阶核，交错鉴定法选择的 h 的阶为 $O_p(n^{-1/(q+4)})$。如果 $q\leqslant3$，则此阶满足条件7.1（iv）、7.2、7.3和7.4。因此，当 $q\leqslant3$ 时，人们可通过最小化 $\sum_{i=1}^{n}[Y_i - X_i'\beta - \hat{g}(Z_i,h)]^2$ 同时选择 h_s 和 β。$\hat{h}_s$ 的阶为 $O_p(n^{-1/(q+4)})$，且$\hat{\beta}$是$\sqrt{n}$一致的，其渐近方差在定理7.1中给出。人们也可利用一个部分线性回归模型中的二阶展开结果（在 h_s 处）选择平滑参数来最小化直到二阶的均方误差的估计值（见Linton（1995））。

局部线性估计量被应用于一系列的设定中。例如，Anglin 和 Gencay（1996）将其应用于特征价格函数的半参数模型；Blundell，Duncan 和 Pendakur（1998）考

虑了恩格尔曲线半参数估计的部分线性模型;Engle, Granger, Rvice 和 Weiss(1986)使用部分线性设定来估计天气与电力销售的关系;Stock(1989)在部分线性框架下考虑了某些政策相关变量分布的变化对被解释变量的平均效应的预测问题;Adams, Berger 和 Sickles(1999)应用一个部分线性设定来研究美国银行业的产量边界;Yatchew 和 No(2001)在一个部分线性模型中让价格和年龄非参数地进入,在允许价格内生性情况下估计了价格和收入弹性。

230

7.3 Andrews 的 MINPIN 方法

Andrews(1994)提供了一个一般框架来证明更广泛一类的半参数估计量的 $\sqrt{n}$—致性和渐近正态性。Andrews 将该估计量命名为 MINPIN,因为它是通过最小化一个准则函数得到的估计量,该准则函数依赖于初始无限维的冗余参数(proliminary infinite dimentional nuisance parameter)的估计量。Andrews 的方法可用于推导许多半参数估计量的渐近分布,包括在一个部分线性模型中 β 的估计量。我们在下文简单地讨论此方法。

令 $\theta \in \Theta \subset \mathbb{R}^p$ 表示一个有限维参数,并令 τ 表示一些无限维的函数。进而,令 $\hat{\tau}$ 为 $\tau \in \mathcal{H}$ 的初始的非参数估计量,其中 $\mathcal{H}$ 是一类平滑函数,该平滑函数的特征依赖于所考虑的特定的半参数模型。我们用 θ_0 和 τ_0 分别代表真实参数和真实的未知函数。令 $\hat{\theta}$ 表示 θ_0 的估计量,它是一个函数最小化问题的解,其中目标函数依赖 θ 和 $\hat{\tau}$,$\hat{\tau}$ 是 τ_0 的初始非参数估计量。假定 $\hat{\theta}$ 为 θ_0 的一致估计量,通过求解一个最小化问题得到如下一阶条件:

$$\sqrt{n}\,\bar{m}_n(\theta,\hat{\tau}) = 0$$

其中 $\bar{m}_n(\theta,\hat{\tau}) = n^{-1}\sum_i m(W_i,\theta,\hat{\tau})$。

我们考虑 $m(W_i,\theta,\hat{\tau})$ 是关于 θ 可微的情形。如果 τ 是有限维的,通过在 (θ_0,τ_0) 处展开 $\sqrt{n}\bar{m}_n(\hat{\theta},\hat{\tau})$,使用元素对元素均值展开式,我们可以建立 $\hat{\theta}$ 的渐近正态性。然而,因为 τ 是无限维的,在 (θ,τ) 的均值展开便不可行。Andrews(1994)建议仅仅对 $\sqrt{n}\bar{m}_n(\hat{\theta},\hat{\tau})$ 关于 θ_0 进行展开,并利用随机等价连续性(stochastic equicontinuity)(见附录 A)概念来处理 $\hat{\tau}$。一个 $\hat{\theta}$ 关于 θ_0 的均值展开可得到

231

$$o_p(1) = \sqrt{n}\bar{m}_n(\hat{\theta},\hat{\tau}) = \sqrt{n}\bar{m}_n(\theta_0,\hat{\tau}) + \frac{\partial}{\partial\theta'}\bar{m}_n(\bar{\theta},\hat{\tau})\sqrt{n}(\hat{\theta}-\theta_0) \quad (7.16)$$

其中 $\bar{\theta}$ 位于 $\hat{\theta}$ 和 θ_0 的线性部分。在一些正则假定下,可以证明

$$\frac{\partial}{\partial\theta'}\bar{m}_n(\bar{\theta},\hat{\tau}) \equiv n^{-1}\sum_i \frac{\partial}{\partial\theta'}m(W_i,\bar{\theta},\hat{\tau})$$

$$\xrightarrow{p} E\left[n^{-1}\sum_i \frac{\partial}{\partial\theta'}m(W_i,\theta_0,\tau_0)\right]$$

$$= E\left[\frac{\partial}{\partial\theta'}m(W_i,\theta_0,\tau_0)\right]$$

$$\equiv M \tag{7.17}$$

于是,假定 M 是非奇异的,从式(7.16)和式(7.17)我们得到

$$\sqrt{n}(\hat{\theta}-\theta_0) = (-M^{-1}+o_p(1))n^{-1/2}\sum_i m(W_i,\theta_0,\hat{\tau}) \tag{7.18}$$

因此,如果

$$n^{-1/2}\sum_i m(W_i,\theta_0,\hat{\tau}) - n^{-1/2}\sum_i m(W_i,\theta_0,\tau_0) = o_p(1) \tag{7.19}$$

那么我们将有

$$\sqrt{n}(\hat{\theta}-\theta_0) = -M^{-1}n^{-1/2}\sum_i m(W_i,\theta_0,\tau_0) + o_p(1)$$

$$\xrightarrow{d} N(0,M^{-1}SM^{-1}) \tag{7.20}$$

根据 Lindeberg-Levy 中心极限定理,其中 $S=\mathrm{var}(m(W_i,\theta_0,\tau_0))$。

实际上,相比简单的部分线性模型,对于更复杂的一般的半参数模型,式(7.19)是很难验证的。Andrews(1994)建议使用"随机等价连续性"概念建立式(7.19)。随机等价连续性可用来建立式(7.19)是因为如果当 $n\to\infty$ 时 $\rho(\hat{\tau},\tau_0)\to 0$($\rho(\cdot,\cdot)$是伪可度量(pseudometric),见附录 A.32 的定义),且 $\nu_n(\tau)$关于 $\tau\in\Lambda$ 是随机等价连续性,其中 Λ 是一个包含内点 τ_0 的有界集,那么

$$|\nu_n(\tau)-\nu(\tau_0)|\xrightarrow{p}0 \tag{7.21}$$

参见 Andrews(1994,式(2.10))对式(7.21)的证明。回忆 $\nu(\tau)\stackrel{\mathrm{def}}{=}\sqrt{n}\bar{m}_n(\theta_0,\tau)$。于是,如果 $\nu_n(\cdot)$具有随机等价连续性,式(7.19)成立。

下面的假定可用来建立关于 $\hat{\theta}$ 的$\sqrt{n}$正态性结果。

232

假定 MINPIN 估计量 $\hat{\theta}$ 解决了关于某个目标函数 $d(\theta,\hat{\tau})$的 $\hat{\theta}=\inf_{\theta\in\Theta}d(\theta,\hat{\tau})$ 的最小化问题,并且因此其一阶条件是$\bar{m}_n(\theta,\hat{\tau})=n^{-1}\sum_i m(W_i,\theta,\hat{\tau})=0$。其总体矩条件是 $E[m(W_i,\theta_0,\tau_0)]=0$。定义 $\nu_n(\tau)=n^{1/2}\bar{m}_n(\theta_0,\tau)$,我们做了如下假定。

假定 7.1　(正态性)假定

(i) $\hat{\theta}\xrightarrow{p}\theta_0\in\Theta$,$\Theta$ 是$\mathbb{R}^r$的一个紧子集。

(ii) $\hat{\tau}\xrightarrow{p}\tau_0\in\mathcal{G}$。

(iii) $\sqrt{n}\nu_n(\tau) \to N(0,S)$。

(iv) $\{\nu_n(\cdot)\}$ 在 τ_0 处是随机等价连续的。

(v) $m(\theta,\tau)$ 在 $\theta \in \Theta$ 上是二次连续可微的，当

$$n^{-1}\sum_i m(W_t,\theta,\tau) \to E[m(W_t,\theta,\tau)]$$

$$n^{-1}\sum_i (\partial/\partial\theta)m(W_t,\theta,\tau) \xrightarrow{p} E[m(W_t,\theta,\tau)]$$

$$n^{-1}\sum_i (\partial/\partial\theta)m(W_t,\theta,\tau) \xrightarrow{p} E[(\partial/\partial\theta)m(W_t,\theta,\tau)]$$

在 $\Theta \times \mathcal{G}$ 上一致（即弱大数定律一致）。

定理 7.3 在假定 7.1 下

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, M^{-1}SM^{-1})$$

证明：上面的定理是 Andrews(1994)的定理 1 的特例，因此被省略了。事实上，Andrews 并没假定独立同分布数据，他允许时间序列数据的弱依赖和不相同分布的数据。我们在第 18 章讨论弱依赖的数据。

实际中，随机等价连续性条件（假定 7.1(iv)）是最难验证的部分，特别是对于高度非线性的半参数模型。Andrews(1994)使用定理 7.3 对一个局部特定非线性模型建立了 $\sqrt{n}$ 一致的结果（也可参见 Ai 和 McFadden(1997)）。我们在 7.5 节讨论关于部分线性模型的假定 7.1(iv)的检验。

233

人们认为以上的假定是指目标函数关于它的变量是平滑的；参见 Chen，Lvinton 和 Keilegom(2003)在准则函数是非平滑时关于半参数模型的估计。

7.4 半参数的效率边界

7.4.1 条件同方差误差情形

在这一节我们讨论式(7.1)的（局部）半参数效率边界，并考虑两种方法。我们首先在误差是条件同方差假定下推导（局部）半参数的下界，即 $E(u_i^2|X_i,Z_i) = E(u_i^2)$，其中 u_i 是正态分布。这种方法用到了参数极大似然估计，且是易于理解的。我们遵循了 Newey(1990b)和 Rilstone(1993)的论证。这里的讨论将是非正式的，不过一种严格的方法可在 Newey(1990b)中找到；同样可参见 Begun，Hall，Huang 和 Wellner(1983)以及 Tripathi 和 Severina(2001)关于半参数模型效率边界的一般处理，还有 Pakes 和 Olley(1995)的相关工作。

234

令 $m(z,\delta)$ 为参数子模型，对于某些 δ_0 有 $m(z,\delta_0) = g(z)$，其中 δ 是一个

$k\times1$ 阶的冗余参数向量($k\geq q$)。该参数子模型的参数是 $\psi=(\beta',\delta')$,且关注的参数向量为 β_0。对于每一个参数子模型,有一个向量值的分函数 l_ψ,人们可以区分关注参数 β_0 和冗余参数 δ 使得 $l_\psi'=(l_\beta',l_\delta')$。半参数边界可被解释为参数子模型协方差矩阵的上确界。定义冗余函数的正切集合$\mathcal{J}$为 l_δ 的所有 $k\times1$ 阶线性组合的均方闭包。令$\mathcal{P}[l_\theta|\mathcal{J}]$代表在$\mathcal{J}$空间上 l_θ 的投射且定义 $l_\theta^*=l_\theta-\mathcal{P}[l_\theta|\mathcal{J}]$。于是模型的半参数下界为$\mathcal{V}_\theta=\{E[l_\theta^*l_\theta^{*\prime}]\}^{-1}$。

参数子模型对应的对数似然函数为:

$$\mathcal{L}(\beta,\delta)=\text{常数}-\frac{(Y-X'\beta_0-g(Z,\delta))^2}{2\sigma^2}$$

在真实模型求值时的得分函数为

$$\begin{bmatrix}l_\beta\\l_\delta\end{bmatrix}\equiv\begin{bmatrix}\dfrac{\partial\mathcal{L}}{\partial\beta}\\\dfrac{\partial\mathcal{L}}{\partial\delta}\end{bmatrix}=\begin{bmatrix}(1/\sigma^2)uX\\-(1/\sigma^2)ug^{(1)}(Z)\end{bmatrix}$$

其中我们用到 $\partial g(z,\delta_0)/\partial z=g^{(1)}(z)$。由于 $g(\eta)$未设定,因此 $g^{(1)}(z)$也不能识别。我们得到正切集合 $\Lambda(\cdot)$,l_δ 的所有 $k\times1$ 阶线性组合的均方闭包,可表示为

$$\Lambda=\{u\Delta(z):E\{[\Delta(Z)]^2\}<\infty\}$$

在 $\Lambda(\cdot)$上的 l_β 的投射为$(1/\sigma^2)uE(X|Z)$,因此,我们得到的效率得分为

$$l_\beta^*\stackrel{\text{def}}{=}l_\beta-E(l_\beta\mid\Lambda(\cdot))=(1/\sigma^2)u(X-E(X\mid Z))$$

因此,式(7.1)的半参数效率边界为

$$\mathcal{V}_\beta=\{E[l_\beta^*l_\beta^{*\prime}]\}^{-1}=\sigma^2\{E[(X-\xi)(X-\xi)']\}^{-1}\tag{7.22}$$

其中 $\xi=E(X|Z)$。

于是,在式(7.9)中定义的 $\hat{\beta}$ 在渐近协方差矩阵Σ等于部分线性模型的半参数下界$\mathcal{V}_\beta$的意义上是半参数有效的。

当式(7.1)是一个线性回归模型,即 $g(z)=\alpha+z'\gamma$,α 是一个标量且 $\gamma\in\mathbb{R}^q$ 时,我们也可以将$\mathcal{V}_\beta$与 β_0 最小二乘估计量的渐近方差进行比较,

$$y_i=X_i'\beta_0+\alpha+Z_i'\gamma+u_i\tag{7.23}$$

令 $\hat{\beta}_{\text{ols}}$表示基于式(7.23)的 β_0 的普通最小二乘估计量,易证明

$$\sqrt{n}(\hat{\beta}_{\text{ols}}-\beta_0)\xrightarrow{d}N(0,V_{\text{ols}})\tag{7.24}$$

其中 $V_{\text{ols}}=\sigma^2E\{[X_i-a-bZ_i][X_i-a-bZ_i]'\}^{-1}$(见习题 7.6),其中 a 和 b 分别为 $p\times1$ 与 $p\times q$ 维常数矩阵,$a+bZ_i$ 是 X_i 在关于 Z_i 的线性空间上的最佳线性投射。现在比较 V_R 与 V_{ols},我们知道若 $E(X_i|Z_i)=a+bZ_i$ 是 Z_i 的线性函数,　235

那么有 $V_R = V_{ols}$。然而，若 $E(X_i|Z_i)$ 不是 Z_i 的线性函数，那么，对任一函数 $s(Z_i)$，$E\{[X_i - E(X_i|Z_i)][X_i - E(X_i|Z_i)]'\} - E\{[X_i - s(Z_i)][X_i - s(Z_i)]'\}$ 是半负定的(根据定理 A.3)。于是，当 $E(X_i|z)$ 关于 z 非线性时，$V_R - V_{ols}$ 是正定的。在后面这种情形，半参数有效估计量 $\hat{\beta}_R$ 与使用了额外信息 $g(z) = \alpha + z'\gamma$ 的参数估计量 $\hat{\beta}_{ols}$ 相比是渐近弱有效的。

7.4.2 条件异方差误差情形

一般的条件异方差误差情形下，对一个部分线性模型的半参数有效边界的推导可以参考 Chamberlain (1992)，而 Ai 和 Chen (2003) 考虑了一般的半参数模型的有效估计，部分线性模型作为一种特殊情形被包括其中。下述的讨论基于 Ai 和 Chen。在本节给出的结果中包含视为一个特殊情形的条件同方差误差模型。

考虑部分线性模型

$$Y_i = X_i'\beta_0 + g(Z_i) + u_i, \quad i = 1, \cdots, n \tag{7.25}$$

其中 $E(u_i^2|X_i, Z_i) = \sigma^2(X_i, Z_i)$ 的形式未知。

为了推导出条件异方差误差情形下 β_0 的半参数有效边界，我们首先假定 $\sigma^2(X_i, Z_i)$ 是已知的，然后再着手讨论当 $\sigma^2(x, z)$ 未知时，如何得到一个可行的半参数有效估计量。Ai 和 Chen (2003) 证明了可以通过同时(联合地)关于 β_0 和未知函数 g 最小化如下目标函数来有效地估计 β_0：

$$\inf_{\beta \in \mathcal{B}, g \in \mathcal{G}} E\{[Y_i - X_i'\beta_0 - g(Z_i)]^2 / \sigma^2(X_i, Z_i)\} \tag{7.26}$$

其中 $\mathcal{B}$ 为 $\mathbb{R}^r$ 的紧子集，$\mathcal{G}$ 为一平滑函数族。

在实际应用中，人们使用的是样本均值而不是总体均值(7.26)，从而最小化下式：

$$\inf_{\beta \in \mathcal{B}, g \in \mathcal{G}} \sum_{i=1}^{n} [Y_i - X_i'\beta_0 - g(Z_i)]^2 / \sigma^2(X_i, Z_i) \tag{7.27}$$

最小化过程可以首先处理未知函数 $g(\cdot)$。也就是说，我们首先视 β_0 为一固定常数，对式(7.26)关于 $g(\cdot)$ 进行积分变换，我们得到

236

$$2E\{E[(Y_i - X_i'\beta_0 + g(Z_i))/\sigma^2(X_i, Z_i) \mid Z_i]\delta(Z_i)\} = 0 \tag{7.28}$$

其中 $\delta(Z_i)$ 是 Z_i 的一个任意函数。对式(7.28)解出 $g(Z_i)$ 得到

$$g(Z_i) = \frac{1}{E\left(\frac{1}{\sigma_i^2} \middle| Z_i\right)}\left[E\left(\frac{Y_i}{\sigma_i^2} \middle| Z_i\right) - E\left(\frac{X_i'}{\sigma_i^2} \middle| Z_i\right)\beta_0\right] \tag{7.29}$$

其中 $\sigma_i^2 = \sigma^2(X_i, Z_i)$。

将式(7.29)代入式(7.27)中，我们得到

$$\sum_i \left\{ Y_i - \frac{E\left(\frac{Y_i}{\sigma_i^2} \middle| Z_i\right)}{E\left(\frac{1}{\sigma_i^2} \middle| Z_i\right)} - \left[X_i - \frac{E\left(\frac{X_i}{\sigma_i^2} \middle| Z_i\right)}{E\left(\frac{1}{\sigma_i^2} \middle| Z_i\right)} \right]' \beta \right\}^2 \Bigg/ \sigma_i^2 \tag{7.30}$$

使用缩写符号 $\mathcal{Y}_i = Y_i - E(Y_i/\sigma_i^2 \mid Z_i)/E(1/\sigma_i^2 \mid Z_i)$ 和 $\mathcal{X}_i = X_i - E(X_i/\sigma_i^2 \mid Z_i)/E(1/\sigma_i^2 \mid Z_i)$，式(7.30)变为 $\sum_i [\mathcal{Y}_i - \mathcal{X}_i'\beta_0]^2/\sigma_i^2$。关于 β_0 最小化此目标函数给出 β_0 的一个(不可行的)有效估计量：

$$\tilde{\beta}_{\text{eff}} = \left[\sum_i \mathcal{X}_i\mathcal{X}_i'/\sigma_i^2\right]^{-1} \sum_i \mathcal{X}_i\mathcal{Y}_i/\sigma_i^2$$

$$= \beta + \left[\sum_i \mathcal{X}_i\mathcal{X}_i'/\sigma_i^2\right]^{-1} \sum_i \mathcal{X}_i u_i/\sigma_i^2$$

由标准的大数定律以及 Lindeberg-Levy 中心极限定理可以得到

$$\sqrt{n}(\hat{\beta}_{\text{eff}} - \beta_0) \to N(0, V_0^{-1}) \text{ 依分布收敛} \tag{7.31}$$

其中

$$V_0 = E[\mathcal{X}_i\mathcal{X}_i'/\sigma_i^2]$$

$$\equiv E\left\{\left[X_i - \frac{E\left(\frac{X_i}{\sigma_i^2} \middle| Z_i\right)}{E\left(\frac{1}{\sigma_i^2} \middle| Z_i\right)}\right]\left[X_i - \frac{E\left(\frac{X_i}{\sigma_i^2} \middle| Z_i\right)}{E\left(\frac{1}{\sigma_i^2} \middle| Z_i\right)}\right]' \Bigg/ \sigma_i^2\right\} \tag{7.32}$$

式(7.32)正是 Chamberlain (1992, p. 569, 式(1.9)给出的半参数有效边界。

237

当 $\sigma^2(X_i, Z_i) = \sigma^2(Z_i)$ 时，式(7.32)可以简化为

$$V_{0,R} = E\{[X_i - E(X_i \mid Z_i)][X_i - E(X_i \mid Z_i)]'/\sigma^2(Z_i)\} \tag{7.33}$$

当 $\sigma_i^2 = \sigma^2$ 为一常数时，式(7.32)进一步缩减为定理7.1中的渐近方差 $\sqrt{n}(\hat{\beta} - \beta_0)$。

上述的估计量 $\tilde{\beta}_{\text{eff}}$ 并不是可行的。β_0 的一个可行的有效估计量可以通过将未知条件期望用它们各自的核估计量来替换从 $\hat{\beta}_{\text{eff}}$ 得到，如用 $\hat{E}[Y_i/\hat{\sigma}_j^2 \mid Z_i] = \sum_j Y_j/\hat{\sigma}_j^2 K_{ij} \Big/ \sum_j K_{ij}$ 替换 $E[Y_i/\sigma_j^2 \mid Z_i]$，其中 $K_{ij} = K((Z_i - Z_j)/h)$，进一步用 $\hat{\sigma}_i^2 = \sum_j \hat{u}_j^2 \bar{K}_{h,ij} \Big/ \sum_j \bar{K}_{h,ij}$ 替换 σ_i^2，其中 $\bar{K}_{h,ij} = K((X_i - X_j)/h_x)K((Z_i - Z_j)/h_z)$，而 $\hat{u}_i = \hat{E}(Y_i \mid Z_i) - \hat{E}(X_i \mid Z)'\hat{\beta}$ 为习题7.1中定义的 u_i 的一个一致估计量。要使得可行的估计量确实具有与非可行有效估计量 $\tilde{\beta}_{\text{eff}}$ 相同的渐近分布，某些额外的正则条件是必需的，例如密度函数 $f(x,z)$ 在其支撑集上是非零有界的，或者

使用修剪函数。我们需要指出的是,β_0 的有效半参数估计量是非常复杂的。它需要估计一个 $p+q$ 维((X_i, Z_i) 的维数)的非参数模型,而估计 β_0 与 $g(\cdot)$ 只涉及了 q 维的非参数估计。于是,当误差为条件异方差情形时,“维数诅咒”阻止了研究者对部分线性模型应用有效估计方法。

本章我们的讨论只覆盖了(Y,X,Z)均是无误差地观测的情形。对于存在变量误差(errors-in-variables)的半参数部分线性模型的估计,见 Liang, Härdle 和 Carroll (1999)以及 Liang 和 Wang (2004)。

238

7.5 证 明

7.5.1 定理 7.2 的证明

我们首先给出 $\hat{\beta}_f$ 渐近方差的一个一致估计量,可以证明:

$$\hat{\Phi}_f = \frac{1}{n}\sum_i (X_i - \hat{X}_i)(X_i - \hat{X}_i)'\hat{f}_i^2$$

$$\hat{\Psi}_f = \frac{1}{n}\sum_i [(\tilde{u}_i\hat{f}_i)^2 (X_i - \hat{X}_i)(X_i - \hat{X}_i)'\hat{f}_i^2] \tag{7.34}$$

分别是 Φ 和 Ψ 的一致估计量,这里$\tilde{\mu}_i \equiv (y_i - \hat{y}_i) - (X_i - \hat{X}_i)'\hat{\beta}_f$ 是 μ_i 的一个一致估计量。

自始至终,我们用 $\sum_i$ 表示$\sum_{i=1}^{n}$,用$\sum_{j\neq i}$ 表示 $\sum_{j=1, j\neq i}^{n}$。

由 $Y_i = X_i'\beta_0 + g_i + \mu_i$,我们可得到 $\hat{Y}_i = \hat{X}_i'\beta_0 + \hat{g}_i + \hat{\mu}_i$,其中 $\hat{A}_i = n^{-1}\sum_{j\neq i} A_j K_{h,ij}/\hat{f}_i (A = Y, X, g, u)$。定义 $S_{A\hat{f},B\hat{f}} = n^{-1}\sum_i A_i\hat{f}_i B_i'\hat{f}_i$,$S_{A\hat{f},B\hat{f}} = S_{A\hat{f}}$。使用 $Y_i - \hat{Y}_i = (X_i - \hat{X}_i)\beta + g_i - \hat{g}_i + u_i - \hat{u}_i$,我们得到

$$\hat{\beta}_f = S^{-1}_{(x-\hat{x})\hat{f}} S_{(x-\hat{x})\hat{f},(Y-\hat{Y})\hat{f}} = \beta_0 + S^{-1}_{(x-\hat{x})\hat{f}} S_{(x-\hat{x})\hat{f},(g-\hat{g}+u-\hat{u})\hat{f}} \tag{7.35}$$

因此,我们有

$$\sqrt{n}(\hat{\beta}_f - \beta_0) = S^{-1}_{(x-\hat{x})\hat{f}} \sqrt{n} S_{(x-\hat{x})\hat{f},(g-\hat{g}+u-\hat{u})\hat{f}} \tag{7.36}$$

在下面的命题 7.1 到 7.4,我们证明:

(i) $\sqrt{n}S_{(x-\hat{x})\hat{f},(g-\hat{g})\hat{f}} = o_p(1)$

(ii) $S_{(x-\hat{x})\hat{f}} \xrightarrow{p} \Phi_f$

(iii) $\sqrt{n}S_{(x-\hat{x})\hat{f},\hat{\mu}\hat{f}} = o_p(1)$

(iv) $\sqrt{n}S_{(x-\hat{x})\hat{f},u\hat{f}} = \sqrt{n}S_{vf,uf} + o_p(1) \xrightarrow{d} N(0, \Psi_f)$

239

使用(i)—(iv),从式(7.36)我们可得到:

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_f - \beta_0) &= S^{-1}_{(x-\hat{x})\hat{f}} \sqrt{n} S_{(x-\hat{x})\hat{f},(g-\hat{g}+u-\hat{u})\hat{f}} \\
&= S^{-1}_{(x-\hat{x})\hat{f}} \sqrt{n} \{ S_{(x-\hat{x})\hat{f},(g-\hat{g})\hat{f}} + S_{(x-\hat{x})\hat{f},u\hat{f}} - S_{(x-\hat{x})\hat{f},\hat{u}\hat{f}} \} \\
&= [\Phi_f + o_p(1)]^{-1} \{ o_p(1) + [\sqrt{n} S_{vf,uf} + o_p(1)] + o_p(1) \} \\
&\xrightarrow{d} \Phi_f^{-1} N(0, \Psi_f) \\
&= N(0, \Phi_f^{-1} \Psi_f \Phi_f^{-1})
\end{aligned}$$

这就完成了定理 7.2 的证明。 □

在下面给出的引理中,我们假定(X_i, Z_i)的支撑是有界集来简化证明。对$g_i = g(Z_i)$的所有结果也对$\xi_i = \xi(Z_i) \equiv E(X_i | Z_i)$成立。令$\varepsilon = v_i$或$u_i$,那么$g(Z_i)$、$\xi(x_i, z_i)$和$\sigma_\varepsilon^2(X_i, Z_i) = E[\varepsilon_i^2 | X_i, Z_i]$都是直到$\nu$阶导数有界的有界函数,尽管$u_i$的支撑不需要是有界的。

我们下面将使用缩写记号$E_1(A) = E(A | X_i, Z_i)$,$K_{ij} = K((Z_i - Z_j)/h)$。

如果我们用ξ来替代g,那么下面的引理 7.1 到 7.5 成立。g和ξ之间唯一的不同是g是一个标量函数而ξ是$r \times 1$维。下面的证明对于ξ的每一个元素仍成立,因此对于向量函数ξ成立,因为r是有限的。

同样,为了记号的简洁,我们假定$h_1 = \cdots = h_q = h$。另外,我们可以把$O(h^2)$的结果解释为$O\left(\sum_{s=1}^{q} h_s^2\right)$,把$O(h^q)$解释为$O(h_1 \cdots h_q)$,来得到我们没有施加所有的$h_s$是相等的条件的情形下对应的结果。

同样,为了简化证明,我们在下面的证明中对$\hat{X}_i$和$\hat{Y}_i$使用去一估计量。注意,不使用去一核估计量定理 7.2 仍然有效。

引理 7.1 令$m_i = g(Z_i)$或$m_i = \xi(Z_i)$,那么

$$E_1[(m_i - m_1) K_{h,i1}] = O(h^\nu)$$

证明:注意$g(z)$有有界导数。用 Talor 展开和变量替换可得。 □

引理 7.2

(i) $S_{(\hat{m}-m)\hat{f}} = O_p(h^{2\nu} + h^2 (nh^q)^{-1})$ ($m_i = g_i$或$m_i = \xi_i$)。

(ii) 令$\varepsilon_i = u_i$或v_i,有$S_{\hat{\varepsilon}\hat{f}} = O_p((nh^q)^{-1})$。

240

(i)的证明(我们忽略$(n-1)^{-1}$和n^{-1}的区别):

$$\begin{aligned}
E[|S_{(\hat{m}-m)\hat{f}}|] &= n^{-1} \sum_i E[(\hat{m}_i - m_i)^2 \hat{f}_i^2] \\
&= E[(\hat{m}_1 - m_1)^2 \hat{f}_i^2] \\
&= n^{-2} \sum_{i \neq 1} \sum_{j \neq 1} E[(m_i - m_1) K_{h,i1} (m_j - m_1) K_{h,j1}]
\end{aligned}$$

$$= n^{-2}\sum_{i\neq 1}\Big\{E[(m_i - m_1)^2K_{h,i1}^2]$$

$$+\sum_{j\neq 1,j\neq i}E\{E_1[(m_i - m_1)K_{h,i1}]\times E_1[(m_j - m_1)K_{h,j1}]\}\Big\}$$

$$= n^{-1}\{O(h^2h^{-q}) + nO(h^{2\nu})\}$$

$$= O((nh^{q-2})^{-1} + h^{2\nu})$$

根据引理 7.1。 □

(ii)的证明：

$$E[|S_{\hat{\varepsilon}\hat{f}}|] = n^{-1}\sum_i E[\hat{\varepsilon}_i^2\hat{f}_i^2]$$

$$= E[\hat{\varepsilon}_1^2\hat{f}_1^2]$$

$$= n^{-2}\sum_{i\neq 1}\sum_{j\neq 1}E[\varepsilon_i\varepsilon_jK_{h,i1}K_{h,j1}]$$

$$= n^{-2}\sum_{i\neq 1}E[\varepsilon_i^2K_{h,i1}^2]$$

$$= n^{-1}E[\sigma_\varepsilon^2(X_i,Z_i)K_{h,i1}^2]$$

$$\leqslant Cn^{-1}E(K_{h,i1}^2)$$

$$= O((nh^q) - 1)$$

□

引理 7.3 $S_{(\hat{m}-m)\hat{f},\varepsilon\hat{f}} = o_p(n^{-1/2})(m_i = g_i \text{ 或 } m_i = \xi_i)$。

241 证明：

$$S_{(\hat{m}-m)\hat{f},\varepsilon\hat{f}} = S_{(\hat{m}-m)\hat{f},\varepsilon f} + S_{(\hat{m}-m)\hat{f},\varepsilon(\hat{f}-f)}$$

$$= S_{(\hat{m}-m)\hat{f},\varepsilon f} + (\text{s.o.})$$

$$E\{[S_{(\hat{m}-m)\hat{f},\varepsilon f}]^2\} = n^{-2}\sum_i E[(\hat{m}_i - m_i)^2\hat{f}_i^2\varepsilon_i^2f_i^2]$$

$$= n^{-1}E[(\hat{m}_1 - m_1)^2\hat{f}_1^2\sigma_\varepsilon^2(X_1,Z_1)f_1^2]$$

$$\leqslant Cn^{-1}E[(\hat{m}_1 - m_1)^2\hat{f}_1^2]$$

$$= Cn^{-1}E[S_{(\hat{m}-m)\hat{f}}]$$

$$= n^{-1}o(1)$$

$$= o(1)$$

根据引理 7.2。

因此，

$$S_{(\hat{m}-m)\hat{f},\varepsilon\hat{f}} = S_{(\hat{m}-m)\hat{f},\varepsilon f} + (\text{s.o.}) = o_p(n^{-1/2})$$

□

引理 7.4

$$S_{(\hat{m}-m)\hat{f},\hat{\varepsilon}\hat{f}} = O_p(h(nh^q)^{-1} + h^\nu(nh^q)^{-1/2}) = o_p(n^{-1/2})$$

证明：由 Cauchy 不等式可得：

$$
\begin{aligned}
|S_{(\hat{m}-m)\hat{f},\hat{\varepsilon}f}| &\leqslant \{|S_{(\hat{m}-m)\hat{f}}||S_{\hat{\varepsilon}\hat{f}}|\}^{1/2}\\
&= \{O_p(h^2(nh^q)^{-1}+h^{2\nu})O_p((nh^q)^{-1})\}^{1/2}\\
&= O_p(h(nh^q)^{-1}+h^{\nu}(nh^q)^{-1/2})\\
&= o_p(n^{-1/2})
\end{aligned}
\qquad \square
$$

引理 7.5

(i) $S_{u\hat{f},\hat{v}\hat{f}}=O_p((nh^{q/2})^{-1})$

(ii) $S_{v\hat{f},\hat{u}\hat{f}}=O_p((nh^{q/2})^{-1})$

(iii) $S_{v\hat{f},\hat{v}\hat{f}}=O_p(n^{-1}h^{q/2})$

242

(iv) $S_{\hat{u}\hat{f},\hat{v}\hat{f}}=O_p((nh^q)^{-1})$

(v) $S_{\hat{v}\hat{f}}=o_p(1)$

(ii)和(iii)的证明同(i)。

(i)的证明：

$$
\begin{aligned}
S_{u\hat{f},\hat{v}\hat{f}} &= S_{uf,\hat{v}\hat{f}}+S_{u(\hat{f}-f),\hat{v}\hat{f}}\\
&= S_{uf,\hat{v}\hat{f}}+(\text{s.o.})\\
E\{[S_{uf,\hat{v}\hat{f}}]^2\} &= n^{-2}\sum_i E[u_i^2f_i^2|\hat{v}_i\hat{f}_i|^2]\\
&= n^{-1}E[\sigma^2(X_1,Z_1)f_1^2|\hat{v}_1\hat{f}_1|^2]\\
&\leqslant Cn^{-1}E[|\hat{v}_1\hat{f}_1|^2]\\
&= O(n^{-1}(nh^q)^{-1})\\
&= O((n^2h^q)^{-1})
\end{aligned}
$$

根据引理 7.2。

因此，

$$
S_{uf,\hat{v}\hat{f}}=O((nh^{q/2})^{-1}) \qquad \square
$$

(ii)的证明同(i)。

(iii)的证明同(i)。

(iv)的证明：根据 Cauchy 不等式：

$$
\begin{aligned}
|S_{\hat{u}\hat{f},\hat{v}\hat{f}}| &\leqslant \{|S_{\hat{u}\hat{f}}||S_{\hat{v}\hat{f}}|\}^{1/2}=\{O_p((nh^q)^{-1})O_p((nh^q)^{-1})\}^{1/2}\\
&= O_p((nh^q)^{-1})=o_p(n^{-1/2})
\end{aligned}
$$

(v)的证明由引理 7.2 得到。

命题 7.1　$S_{(X-\hat{X})\hat{f},(g-\hat{g})\hat{f}}=o_p(n^{-1/2})$。

证明：定义 $\xi_i=E(X_i|Z_i)$，$V_i=X_i-E(X_i|Z_i)=X_i-\xi_i$。那么 $X_i=\xi_i+V_i$，$\hat{X}_i=\hat{\xi}_i+\hat{V}_i$。

243

$$
S_{(X-\hat{X})\hat{f},(g-\hat{g})\hat{f}}=S_{(\xi-\hat{\xi}+V-\hat{V})\hat{f},(g-\hat{g})\hat{f}}
$$

$$= S_{(\xi-\hat{\xi})\hat{f},(g-\hat{g})\hat{f}} + S_{V\hat{f},(g-\hat{g})\hat{f}} - S_{\hat{V}\hat{f},(g-\hat{g})\hat{f}}$$
$$= O_p(h^2(nh^q)^{-1} + h^{2\nu}) + O_p(h(nh^{q/2})^{-1} + n^{-1/2}h^{\nu})$$
$$= o_p(n^{-1/2})$$

根据引理 7.2、7.3 和 7.4。 □

命题 7.2 $S_{(X-\hat{X})\hat{f}} \xrightarrow{p} \Phi$。

证明：

$$S_{(X-\hat{X})\hat{f}} = S_{(\xi-\hat{\xi})\hat{f}+V\hat{f}-\hat{V}\hat{f}}$$
$$= S_{(\xi-\hat{\xi})\hat{f}} + S_{V\hat{f}} + S_{\hat{V}\hat{f}} + 2S_{(\xi-\hat{\xi})\hat{f},V\hat{f}} - 2S_{(\xi-\hat{\xi})\hat{f},\hat{V}\hat{f}} - 2S_{V\hat{f},\hat{V}\hat{f}}$$
$$= S_{V\hat{f}} + o_p(1)$$
$$= S_{Vf} + o_p(1)$$
$$= n^{-1}\sum_{i=1}^{n} V_i V_i' f_i^2 + o_p(1) \xrightarrow{p} E(V_1 V_1' f_1^2) = \Phi_f$$

根据引理 7.1 至 7.5 和一个大数定律结论。 □

命题 7.3 $S_{(X-\hat{X})\hat{f},\hat{U}\hat{f}} = o_p(n^{-1/2})$。

证明：

$$S_{(X-\hat{X})\hat{f},\hat{U}\hat{f}} = S_{(\xi-\hat{\xi})\hat{f},\hat{U}\hat{f}} + S_{V\hat{f},\hat{U}\hat{f}} - S_{\hat{V}\hat{f},\hat{U}\hat{f}}$$
$$= O_p(h(nh^q)^{-1} + h^{\nu}(nh^q)^{-1/2})$$
$$+ O_p((nh^q)^{-1}) + O_p((nh^q)^{-1})$$
$$= o_p(n^{-1/2})$$

根据引理 7.4 和 7.5。 □

命题 7.4 $\sqrt{n}S_{(X-\hat{X})\hat{f},U\hat{f}} \xrightarrow{d} N(0,\Psi)$。

证明：

$$\sqrt{n}S_{(\hat{X}-X)\hat{f},U\hat{f}} = \sqrt{n}\{S_{(\xi-\hat{\xi})\hat{f},U\hat{f}} + S_{V\hat{f},U\hat{f}} - S_{\hat{V}\hat{f},U\hat{f}}\}$$
$$= \sqrt{n}S_{V\hat{f},U\hat{f}} + o_p(1)$$
$$= \sqrt{n}S_{Vf,Uf} + o_p(1)$$
$$= n^{-1/2}\sum_i V_i U_i f_i^2 \xrightarrow{d} N(0,\Psi_f)$$

根据引理 7.3、7.5 和 Lindeberg-Levy 中心极限定理。 □

7.5.2 对部分线性模型验证定理 7.3

我们将考虑一个密度加权估计量 $\hat{\beta}$，目标函数为

$$d(\theta,\hat{\tau}) = (1/2)\sum_i [Y_i - \hat{Y}_i - (X_i - \hat{X}_i)'\beta_0]^2 \hat{f}_i^2$$

必需的一阶条件是

$$\bar{m}_n(\beta,\hat{\tau}) = n^{-1}\sum_i (X_i - \hat{X}_i)'\hat{f}_i[Y_i - \hat{Y}_i - (X_i - \hat{X}_i)'\beta_0]\hat{f}_i = 0 \quad (7.37)$$

和

$$Y_i - \hat{Y}_i = (X_i - \hat{X}_i)'\beta_0 + (g_i - \hat{g}_i) + u_i - \hat{u}_i \quad (7.38)$$

使用式(7.38)，式(7.37)可重写成

$$\frac{1}{n}\sum_i (X_i - \hat{X}_i)'\hat{f}_i[g_i - \hat{g}_i + u_i - \hat{u}_i]\hat{f}_i - \frac{1}{n}\sum_i (X_i - \hat{X}_i)(X_i - \hat{X}_i)'(\hat{\beta} - \beta_0) = 0 \quad (7.39)$$

比较式(7.39)和式(7.17)，我们知道

$$\nu_n(\hat{\tau}) = n^{1/2}S_{(\xi-\hat{\xi}+v-\hat{v})\hat{f},(g-\hat{g}+u-\hat{u})\hat{f}}$$

注意这里用 $\hat{\xi}$、$\hat{g}$、$\hat{v}$、$\hat{u}$ 和 $\hat{f}$ 分别估计 ξ、$\hat{g}$、$E(v|z)=0$、$E(u|z)=0$ 和 f。因此，在 $v_n(\hat{\tau})$ 中用 τ_0 取代 $\hat{\tau}$，我们得到 $\nu_n(\hat{\tau}_0) = n^{1/2}S_{vf,uf}$。

245

假定 7.1(v)被满足，因为

$$n^{-1}\sum_i (\partial/\partial\beta_0)m(W_i,\beta_0,\hat{\tau}) = n^{-1}\sum_i (X_i - \hat{X}_i)(X_i - \hat{X}_i)'\hat{f}_i^2 \xrightarrow{p} E[(X_i - E(X_i \mid Z_i))(X_i - E(X_i \mid Z_i))'f_i^2] \equiv E[(\partial/\partial\beta_0)m(W_i,\beta_0,\tau_0)]$$

证明类似于命题 7.2。

为验证假定 7.1(iv)，我们有

$$\nu_n(\hat{\tau}) - \nu_n(\tau_0) = n^{1/2}S_{(\xi-\hat{\xi}-\hat{v})\hat{f},(g-\hat{g}+u-\hat{u})\hat{f}} + n^{1/2}S_{v\hat{f},(g-\hat{g}-\hat{u})\hat{f}} \equiv \nu_{n,1(\hat{\tau},\tau_0)} + \nu_{n,2(\hat{\tau},\tau_0)}$$

其中

$$\nu_{n,1(\hat{\tau},\tau_0)} = n^{1/2}S_{(\xi-\hat{\xi}-\hat{v})\hat{f},(g-\hat{g}+u-\hat{u})\hat{f}}$$

$$\nu_{n,2(\hat{\tau},\tau_0)} = n^{1/2}S_{v\hat{f},(g-\hat{g}-\hat{u})\hat{f}}$$

我们定义伪度量：$\rho_n(\hat{\tau},\tau_0) = E[\nu_{n,1}(\hat{\tau},\tau_0)] + E[\nu_{n,2}(\hat{\tau},\tau_0)^2]$，$\rho(\hat{\tau},\tau_0) \stackrel{\text{def}}{\rightarrow} \lim_{n\to\infty}\rho_n(\hat{\tau},\tau_0)$，那么

$$\rho_n(\hat{\tau},\tau_0) = E[\nu_{n,1}(\hat{\tau},\tau_0)] + E[\nu_{n,2}(\hat{\tau},\tau_0)^2] \leqslant E\{n^{-1/2}\sum_i[(\xi_i - \hat{\xi}_i)^2\hat{f}_i^2 + \hat{v}_i^2\hat{f}_i^2 + (g_i - \hat{g}_i)^2\hat{f}_i^2 + \hat{u}_i^2\hat{f}_i^2]\} + n^{-1}\sum_i E\{v_i^2[(g_i - \hat{g}_i)^2\hat{f}_i^2 + \hat{u}_i^2\hat{f}_i^2]\} \to 0$$

根据引理 7.2 的证明。

因此,假定 7.1(iv)成立。

注意对局部线性估计模型 $m(\hat{\theta},\hat{\tau})$ 关于 $\hat{\tau}$ 是线性的。所以随机等连续性容易被证明。随机等连续性的性质实质上对非线性函数族是很难验证的。

由引理 7.3,我们知道

$$\sqrt{n}(\hat{\beta}-\beta_0)\xrightarrow{d}N(0,M^{-1}SM^{-1})$$

这当然和引理 7.1 一样,因为 $M=\Phi_f,\Psi_f=S$。

246

7.6 习 题

习题 7.1 证明式(7.5)。

提示:记 $W_i=X_i-E(X_i|Z_i)$,有

$$\sqrt{n}(\hat{\beta}_{\inf}-\beta_0)=\left[n^{-1}\sum_{i=1}^{n}W_iW_i'\right]^{-1}n^{-1/2}\sum_{s=1}^{n}W_iu_i\equiv A_n^{-1}B_n$$

根据 Khinchin 大数定律(见附录 A),我们有 $A_n=n^{-1}\sum_i W_iW_i\xrightarrow{p}E[W_iW_i']\equiv\Phi$,同时,$E[W_iu_i]=0,\operatorname{var}(W_iu_i)=E[\sigma^2(X_i,Z_i)W_iW_i']\equiv\Psi$。由 Lindeberg-Levy 中心极限定理,我们有

$$n^{-1/2}\sum_i W_iu_i\xrightarrow{d}N(0,\Psi)$$

这一结论意味着

$$\sqrt{n}(\hat{\beta}-\beta_{\inf})=[A_n]^{-1}B_n\to\Phi^{-1}N(0,B)=N(0,\Phi^{-1}B\Phi^{-1})$$

习题 7.2 证明 $\hat{\beta}$ 渐近方差的一个一致估计量由 $\hat{\Phi}^{-1}\hat{\psi}\hat{\Phi}^{-1}$ 给出,其中 $\hat{\Phi}=\frac{1}{n}\sum_i(X_i-\hat{X}_i)'(X_i-\hat{X}_i)\mathbf{1}_i,\hat{\Psi}=\frac{1}{n}\sum_i[\hat{u}_i^2(X_i-\hat{X}_i)'(X_i-\hat{X}_i)\mathbf{1}_i]$,$\hat{u}_i=(y_i-\hat{y}_i)-(X_i-\hat{X}_i)'\hat{\beta}$ 是 u_i 的一个一致估计量,$\hat{y}_i$ 和 $\hat{X}_i$ 是 $E(y_i\mid Z_i)$ 和 $E(X_i\mid Z_i)$ 各自的核估计量,定义见 7.2 节。

习题 7.3 证明 $\hat{\beta}_f$ 渐近方差的一个一致估量由 $\hat{\Phi}_f^{-1}\hat{\psi}_f\hat{\Phi}_f^{-1}$ 给出,其中 $\hat{\Phi}_f=n^{-1}\sum_i(X_i-\hat{X}_i)'(X_i-\hat{X}_i)\hat{f}_i^2,\hat{\psi}_f=n^{-1}\sum_i[\hat{u}_i^2(X_i-\hat{X}_i)'(X_i-\hat{X}_i)\hat{f}_i^4]$,$\tilde{u}_i=(Y_i-\hat{Y}_i)-(X_i-\hat{X}_i)'\hat{\beta}_f$ 是 u_i 的一个一致估计量,$\hat{Y}_i$ 和 $\hat{X}_i$ 是 $E(y_i\mid Z_i)$ 和 $E(X_i\mid Z_i)$ 各自的核估计量,定义见 7.2 节。

习题 7.4 证明在条件 $\hat{\beta}-\beta_0=O_p(n^{-1/2})$ 下,

$$(nh_1\cdots h_q)^{1/2}\left(\hat{g}(z)-g(z)-\sum_{s=1}^{q}h_s^2B_s(z)\right)\to N(0,V(z))\text{ 依分布收敛}$$

其中 $\hat{g}(z)$ 的定义见式(7.13)，$B_s(z)$ 定义同定理 7.1，$V(z)=\kappa^q[\sigma^2(z)/f(z)]$。

提示：根据定理 7.1，$(nh_1\cdots h_q)^{1/2}(\tilde{g}(z)-g(z)-\sum_{s=1}^{q}h_s^2B_s(z))\to N(0,V(z))$，$\tilde{g}(z)$ 的定义见式(7.14)。由 $\hat{\beta}-\beta_0=O_p(n^{-1/2})$，可以证明 $\hat{g}(z)-\tilde{g}(z)=O_p(n^{-1/2})=o_p((nh_1\cdots h_q)^{-1/2}+\sum_{s=1}^{q}h_s^2)$。

习题 7.5

(i) 证明引理 7.4 的结论可以被加强为：

$$S_{(\hat{g}-g)\hat{f},\hat{\varepsilon}\hat{f}}=O_p(h\,(nh^{q/2})^{-1}+n^{-1/2}h^{\nu})=o_p(n^{-1/2})$$

(ii) 证明引理 7.5(iii)的结论可以被加强为：$S_{\hat{u}\hat{f},\hat{v}\hat{f}}=O_p(n^{-1}h^{-q/2})$。

提示：不要使用 Cauchy 不等式，而是对(i)证明：

$$E[S^2_{(\hat{g}-g)\hat{f},\hat{\varepsilon}\hat{f}}]=O(h^2\,(n^{-1}h^q)^{-1}+n^{-1}h^{2\nu})$$

对(ii)证明 $E[S^2_{\hat{u}\hat{f},\hat{v}\hat{f}}]=O(n^{-2}h^{-q})$。

注意在上面的加强中，条件 7.3 可以用较弱的条件 7.4 来替代。

习题 7.6　证明式(7.24)。

提示：利用 Frisch-Waugh-Lovell 定理(见 Davidson 和 MacKinnon (1993, pp.19—24))。定义$\mathcal{M}_z=I_n-P_z(P_z'P_z)^{-1}P_z'$，其中 P_z 是一个 $n\times(q+1)$ 阶矩阵，其第 i 行为$(1,Z_i')$。将式(7.23)的矩阵形式投射到$\mathcal{M}_z$以消除 $\alpha+Z_i'\gamma$，然后使用一个标准的大数定律和中心极限定理，正如我们在习题 7.1 的证明中所做的。

习题 7.7

(i) 假定 $E(u_i^2|X_i,Z_i)=\sigma^2(Z_i)$，使用类似在推导式(7.32)时我们使用过的论证，证明 β_0 的半参数的有效边界由式(7.32)给出。

(ii) 当 $E(u_i^2|X_i,Z_i)=\sigma^2(X_i,Z_i)$ 时，$V_0-V_{0,R}$是正定的吗？

提示：回答(ii)不需要计算，一个简单的逻辑论证就足够了。

第8章 半参数单指标模型

249

本章我们考虑另外一个应用非常广泛的半参数模型,即所谓的半参数单指标模型。该模型已经被计量经济学家广泛运用于各种情形。

半参数单指标模型具有如下形式

$$Y = g(X'\beta_0) + u \tag{8.1}$$

其中 Y 是因变量,$X \in \mathbb{R}^q$ 为解释变量向量,β_0 为未知参数的 $q \times 1$ 向量,u 为误差项,满足 $E(u|X) = 0$。$x'\beta_0$ 项被称为"单指标"是因为它是一个标量,尽管 x 是一个向量。$g(\cdot)$ 的函数形式对研究者是未知的。这个模型是半参数的,本质上是因为该模型线性指标的函数形式是已知的,而 $g(\cdot)$ 的形式是未知的。

半参数单指标模型自然地产生于二元选择设定,出于说明的目的我们首先讨论一个流行的例子。在一个二元选择模型中,如果人们愿意接受参数线性指标来控制选择而不愿意设定误差项的未知分布,这时人们就得到了一个半参数单指标模型。特别地,当考虑一个二元因变量(Y)与其他协变量(X)之间的关系时,这种关系可能被人们模型化为

$$Y_i = \begin{cases} 1, Y_i^* \overset{\text{def}}{=} \alpha + X_i'\beta + \varepsilon_i > 0 \\ 0, Y_i^* = \alpha + X_i'\beta + \varepsilon_i \leqslant 0 \end{cases} \tag{8.2}$$

250

其中 Y^* 是一个潜变量。注意这里 $\varepsilon = Y^* - E(Y^* | X)$,它不同于在式(8.1)中定义的 $u = Y - E(Y|X)$,因为 $Y \neq Y^*$。

例如,Y 值为1时可以表示为劳动力参与决策,即在个体参与劳动力市场中,当 Y 值为0时则表示个体不参与。解释变量 X 包含一组可以影响参与决策的经济因素,如年龄、婚姻状况、受教育程度、工作经历和子女数目等。模型(8.2)假定是否参与劳动力市场的决定 Y 与

解释变量 X 之间是由一个线性参数连接函数联系到一起的。我们主要目的是估计 β，它反映了 X 的变化对劳动力市场参与概率的影响。

估计 β 的参数方法要求给定误差项 ε 的（未知）分布。一个常见的假定是 ε 服从一个正态分布，即 $\varepsilon \sim N(0,\sigma^2)$。可以证明，在没有其他识别条件的情况下，$\beta$ 和 σ^2 是不能被联合识别的（更多细节见 Maddala(1986)）。例如，如果我们假定 $\sigma=1$，那么 β 是可识别的，我们可以用极大似然估计来估计 β。然而，若误差项并不服从正态分布，那么一般来说参数方法将得到不一致的估计，即 $P(Y=1|x)=E(Y|x)$；见习题 8.1。为了认识这一点，令 $F_\varepsilon(\cdot)$ 表示 ε 的真实累积分布函数。由式(8.2)可得

$$
\begin{aligned}
E(Y|x) &= \sum_{y=0,1} yP(y|x) \\
&= 1\times P(Y=1|x)+0\times P(Y=0|x) \\
&= P(Y=1|x) \\
&= P(\alpha+x'\beta+\varepsilon>0) \\
&= P(\varepsilon>-(\alpha+x'\beta)) \\
&= 1-P(\varepsilon<-(\alpha+x'\beta)) \\
&= 1-F(-(\alpha+x'\beta)) \\
&\equiv m(\alpha+x'\beta)
\end{aligned}
$$

其中 $F(\cdot)$ 是 ε 的累积分布函数。注意如果 ε 的分布是对称的，则有 $F(\alpha+x'\beta)=1-F(-(\alpha+x'\beta))$，此种情形下我们有 $m(\cdot)=F(\cdot)$。例如，若 $\varepsilon\sim N(0,1)(\sigma=1)$，那么式(8.2)变成一个 Probit 模型：

$$E(Y|x)=P(Y=1|x)=\Phi(\alpha+x'\beta) \tag{8.3}$$

其中 $\Phi(\cdot)$ 是标准正态变量的累积分布函数。另一方面，如果 ε 服从对称 logistic 分布，那么由式(8.2)可导出 logistic 模型，形式为

251

$$E(Y|x)=P(Y=1|x)=\frac{1}{1+\mathrm{e}^{\alpha+x'\beta}} \tag{8.4}$$

由式(8.3)和式(8.4)，我们可以看出，对于不同的 ε 的分布假定，会得出相异的 $Y=1$ 的条件概率的函数形式。因此，$P(Y=1|x)=E(Y|x)$ 的一致参数估计要求 ε 正确的分布设定。一个半参数单指标模型则避免了有关误差项分布设定错误的问题。此外，半参数单指标模型(8.1)比二元选择模型更一般化，这是因为实际上我们不再要求因变量必须是二元的。正如我们即将在 8.1 节见到的，当 $g(\cdot)$ 的具体函数形式未知时，位置参数 α 是不可识别的，这就是我们把半参数模型(8.1)只写作 $X_i'\beta_0$ 的一个函数的原因。我们会在 8.1 节讨论这个条件和其他的识别条件。注意模型(8.1)意味着 $E(Y|x)=g(x'\beta_0)$，于是，x 只通过线性组合 $x'\beta_0$ 来影响 y，这种关系由连接函数 $g(\cdot)$ 来刻画。

我们在这里要强调的是,与部分线性模型一样,半参数单指标模型是一种可供选择用来减轻维数诅咒影响的方法。另外,我们强调 Y 可以是连续的或者离散的,即没有理由限定 Y 为一个二元变量。

8.1 识别条件

对于半参数单指标模型,我们有

$$E(Y \mid x) = g(x'\beta_0)$$

Ichimura(1993)、Manski(1988)和 Horowitz(1998, pp. 14—20)提供了隐含在半参数单指标模型中的识别条件极好的直观解释(即在这一系列条件下未知参数向量 β_0 和未知函数 $g(\cdot)$ 能被明智地估计出)。我们简要讨论一下这些条件,再把其概括为一个命题。

252

首先,$g(\cdot)$ 不能是常数方程,否则 β_0 显然不能被识别。其次,与线性回归的情形一样,x 的不同组成部分不能具有完全线性关系(完全多重共线性)。另一个限制是 x 至少包含一个连续随机变量。若 x 只含有离散变量,如某些 0-1 虚拟变量,那么 x 的支撑集是有限集,对于任一向量 β,标量 $v = x'\beta$ 的支撑集也是有限集。那么,显然存在无穷多的函数 $g(\cdot)$,它们的不同在于 β 向量,即 $g(x'\beta) = E(Y|x)$。这是因为 $E(Y|x) = g(x'\beta)$ 只给出了关于未知函数 $g(\cdot)$ 的有限个约束,故存在无穷多个 β 和 $g(\cdot)$ 的选择使得它们满足 $E(Y|x) = g(x'\beta)$ 给出的有限个约束。对此的一个详尽的例子和说明见 Horowitz(1998)。同样,x 也不能含有常数,也就是说,β_0 不能包含位置参数,β_0 仅仅在一定规模内是可识别的。这是因为,对任一非零常数 α_1 和 α_2 以及对任一 $g(\cdot)$ 和固定向量 β,我们总能找到另外一个函数,称之为 $g_2(\cdot)$,定义为 $g_2(\alpha_1 + \alpha_2 x'\beta) = g(x'\beta)$,所以,没有位置和规模约束(正规化)$\beta_0$ 不能被识别。一个常用的正规化方法是 x 不含有常数,即所谓的位置归零化(location normalization)。对于所谓的尺度归一化(scale normalization),一种方法是假定向量 β 具有单位长度,也即 $\|\beta\| = 1$,其中 $\|\beta\| = \{\sum_{j=1}^{q}\beta_j^2\}^{1/2}$ 是 β 的欧几里得范数(长度)。另一种方法是假定 x 的第一个元素具有单位系数,而且第一个元素为连续变量。

我们把上述条件概括为以下命题。

命题 8.1 (单指标模型的识别。)对于半参数单指标模型(8.1),识别 β_0 和 $g(\cdot)$ 要求:

(i) x 不含有常数(截距),且 x 必须含有至少一个连续变量。此外,$\|\beta_0\| = 1$。

(ii) $g(\cdot)$ 是可微的且在 $x'\beta_0$ 的支撑集上为非常数函数。

（iii）对于 x 的离散元，改变这些离散变量的赋值不会把 $X'\beta$ 的支撑集划分成分离的子集。

我们已经讨论过当 x 只含有离散变量时，β_0 和 $g(\cdot)$ 是如何不可识别的。 253 然而，当 $g(\cdot)$ 被假定为增函数时，人们可以得到 β 中元素的可识别界限。当 x 的所有元素是离散时，对如何刻画界限的更详尽讨论见 Horowitz(1998，pp. 17—20)。

8.2 估　计

8.2.1 Ichimura 方法

本节我们回顾 Ichimura(1993)提出的半参数估计方法。

若 $g(\cdot)$ 的函数形式是已知的，式(8.1)就变为标准的非线性回归模型，我们就可以用非线性最小二乘方法来估计 β_0，关于 β 最小化

$$\sum_{i=1}[Y_i - g(X_i'\beta)]^2 \tag{8.5}$$

$g(\cdot)$ 函数形式未知的情形下，我们首先要估计 $g(\cdot)$。然而，核方法不能直接估计 $g(X_i'\beta_0)$，因为不仅 $g(\cdot)$ 未知，而且 β_0 也未知。然而，对于一个给定的 β 值我们能通过核方法估计

$$G(X_i'\beta) \stackrel{\text{def}}{=} E(Y_i \mid X_i'\beta) = E[g(X_i'\beta_0) \mid X_i'\beta] \tag{8.6}$$

其中后一个等式是由于 $E(u_i|X_i)=0$，则对所有 β，有 $E(u_i|X_i'\beta)=0$。

注意当 $\beta=\beta_0$ 时，有 $G(X_i'\beta)=g(X_i'\beta_0)$，一般来说，若 $\beta \neq \beta_0$，则 $G(X_i'\beta) \neq g(X_i'\beta_0)$。$G(X_i'\beta)$ 的一个去一非参数核估计量如下

$$\hat{G}_{-i}(X_i'\beta) \equiv E_{-i}(Y_i \mid X_i'\beta) = \frac{(nh)^{-1}\sum_{j=1,j\neq i}^{n} Y_j K\left(\frac{X_j'\beta - X_i'\beta}{h}\right)}{\hat{p}_{-i}(X_i'\beta)} \tag{8.7}$$

其中 $\hat{p}_{-1}(X_i'\beta) = (nh)^{-1}\sum_{j=1,,j\neq i}^{n} K\left(\frac{X_j'\beta - X_i'\beta}{h}\right)$。

Ichimura(1993)建议用式(8.7)中的 $G_{-i}(X'\beta)$ 来估计式(8.5)中的 $g(X_i'\beta)$，用(半参数)非线性最小二乘法来选择 β。然而这里存在一个技术问 254 题，就是式(8.7)中有一个随机分母，即 $\hat{p}_{-i}(X'\beta) = (nh)^{-1}\sum_{j=1,j\neq i}^{n} K\left(\frac{X_j'\beta - X_i'\beta}{h}\right)$。Ichimura 用一个修剪函数来剔除 $\hat{p}_{-i}(X_i\beta)$ 很小的值。令 $p(x'\beta)$ 为 $x_i'\beta$ 的概率密度函数，定义集合 A_δ 和 A_n 如下：

$$A_\delta = \{x : p(x'\beta) \geq \delta, \text{所有 } \beta \in \mathcal{B}\}$$

其中 $\delta>0$ 是一个常数，$\mathcal{B}$是$\mathbb{R}^q$中的一个紧子集。

$$A_n = \{x: \| x - x^* \| \leqslant 2h, 某些\ x^* \in A_\delta\}$$

集合 A_δ 保证了当 $x\in A_\delta$时，式(8.7)中的分母不会距零太近。集合 A_n 比 A_δ 稍大一些，当 $n\to\infty$，$h\to 0$ 时，A_n 收缩为 A_δ。

Ichimura(1993)建议通过最小化下列目标函数来选择β：

$$S_n(\beta) = \sum_{i=1}^{n} [Y_i - \hat{G}_{-i}(X_i'\beta)]^2 w(X_i)\mathbf{1}(X_i \in A_n) \tag{8.8}$$

其中 $\hat{G}_{-i}(X_i'\beta)$是式(8.7)中定义的，$w(X_i)$是一非负的权重函数，$\mathbf{1}(\cdot)$是通常的示性函数。也就是说，$\mathbf{1}(X_i\in A_n)$是一个借势函数，当 $X_i\in A_n$ 时其值为1，反之则为0。

修剪函数保证了核估计量中的随机分母有大概率为正值，这样就简化了渐近分析。

定义$\hat{\beta}$为从最小化式(8.8)中得到的关于β_0的半参数估计量。为了推导$\hat{\beta}$的渐近分布，以下条件是必需的：

假定8.1 集合 A_δ 是紧集，加权函数 $w(\cdot)$在 A_δ 上为正且有界。定义集合 $D_z=\{z:z=x'\beta,\beta\in\mathcal{B},x\in A_\delta\}$，令 $p(\cdot)$为 $z\in D_z$ 的概率密度函数，对所有的 $z\in D_z$，$p(\cdot)$是有界的且小于一个正的常数。

假定8.2 $g(\cdot)$和 $P(\cdot)$关于 $z=x'\beta$ 是三阶可微的。其三阶导数对于所有的 $z\in D_z$，在$\mathcal{B}$上均为 Lipschitz 连续的。

假定8.3 核函数为有界二阶核，具有有界支撑集，二阶可微，且其二阶导数是 Lipschitz 连续的。

255

假定8.4 对某些 $m\geqslant 3$，$E|Y^m|<\infty$。对所有 $x\in A_\delta$，$\mathrm{cov}(Y|x)$有界且远离零。当 $n\to\infty$ 时，有 $q\ln(h)/[nh^{3+3/(m-1)}]\to 0$ 和 $nh^8\to 0$。

Ichimura(1993)证明如下结果：

定理8.1 在假定8.1至8.4的条件下，

$$\sqrt{n}(\hat{\beta}-\beta_0)\to N(0,\Omega_I)\ 依分布收敛$$

其中 $\Omega_I=V^{-1}\Sigma V^{-1}$，且

$$\Sigma = E\{w(X_i)\sigma^2(X_i)(g_i^{(1)})^2(X_i - E_A(X_i|X_i'\beta_0)) \times(X_i - E_A(X_i|X_i'\beta_0))'\}$$

其中 $g_i^{(1)}=[\partial g(v)/\partial v]|_{v=X'\beta_0}$，$E_A(X_i|v)=E(X_i|X'\beta_0=v)$，其中 x_A 具有 X_i 关于 $X_i\varepsilon A_\delta$ 的条件分布，且

$$V = E[w(X_i)(g_i^{(1)})^2(X_i - E_A(X_i|X_i'\beta_0))(X_i - E_A(X_i|X_i'\beta_0))']$$

Ω_I 的一个一致估计量如下

$$\hat{\Omega}_I = \hat{V}^{-1}\hat{\Sigma}\hat{V}^{-1}$$

其中 $\hat{V} = n^{-1}\sum_i w(X_i)(\hat{g}^{(1)}(X_i'\hat{\beta}))^2(X_i - \hat{E}(X_i \mid X_i'\beta))(X_i - \hat{E}(X_i \mid X_i'\beta))'$，$\hat{\Sigma} = n^{-1}\sum_i w(X_i)\hat{u}_i^2(\hat{g}^{(1)}(X_i'\hat{\beta}))^2(X_i - \hat{E}(X_i \mid X_i'\beta))(X_i - \hat{E}(X_i \mid X_i'\beta))'$，$\hat{u}_i = Y_i - \hat{g}(X_i'\hat{\beta})$，$\hat{g}^{(1)}(X_i'\hat{\beta}) = [\partial \hat{g}_{-i}(X_i'\beta)/\partial X_i'\beta] \mid_{\beta=\hat{\beta}}$，$\hat{g}_{-i}(X_i'\beta)$ 如式(8.7)中所定义的，$\hat{E}(X_i \mid X_i'\beta)' = \sum_j X_j K[(X_i - X_j)'\beta] / \sum_j K[(X_i - X_j)'\beta]$。

定理 8.1 的证明非常具有技巧性，我们推荐读者去参考 Ichimura(1993)。Horowitz(1998)给出了证明定理 8.1 的一个漂亮的启发性提纲，仅使用了人所熟知的 Taylor 级数方法、一个标准的大数定律和 Linderberg-Levy 中心极限定理来论证。

当 $E(u_i^2 \mid X_i) = \sigma^2$ 为一个常数(即 $u_i = Y_i - g(X_i'\beta_0)$ 有条件同方差)时，可以证明，$w(X_i)$ 的最优选择为 $w(X_i) = 1$，在这种情况下，在 Ω_I 是半参数方差下界($X \in A_\delta$ 条件下)的意义上，$\hat{\beta}$ 是半参数有效的。

然而，一般来说 $E(u_i^2 \mid X_i) = \sigma^2(X_i)$，且 β 的半参数有效估计量有一个复杂的结构。若假定 $E(u_i^2 \mid X_i) = \sigma^2(X'\beta_0)$，也即条件方差只依赖单指标，那么 $w(X_i) = 1/\sigma^2(X'\beta_0)$ 的选择能导致一个 β_0 的半参数有效估计量。但是，实践中，由于 $\sigma^2(X'\beta_0)$ 是未知的，以上权重函数的选择是不可行的。因此人们可以采取如下的两步方法。假定条件方差只是 $X_i'\beta_0$ 的函数。在这种情形下，第一步，先用 $w(X_i) = 1$ 来得到 β_0 的一个$\sqrt{n}$一致估计量，比如 $\tilde{\beta}$。然后，运用 $\tilde{u}_i = Y_i - \hat{g}(X_i'\tilde{\beta}_0)$ 可以得到 $\sigma^2(X_i'\beta_0)$ 的一个一致的非参数估计量，比如 $\hat{\sigma}^2(X'\tilde{\beta}_0) = \widehat{\text{var}}(\hat{u}_i \mid X_i\tilde{\beta}_0)$。第二步，选择 $w(X_i) = 1/\hat{\sigma}^2(X_i'\tilde{\beta}_0)$ 来再次估计 β_0。如果在 $v \in D_v$(D_v 为 $X'\beta_0$ 的支撑集)上，$\hat{\sigma}^2(v) - \sigma^2(v)$ 以一特殊速率一致收敛于 0，那么相应的两步估计量 $\hat{\beta}_0$ 将是半参数有效的。

256

以下我们忽略修剪集 A_δ、权重函数 $w(\cdot)$，并同样假定关于 β 最小化，事实上是在一个收缩集 $\mathcal{B}_n = \{\beta: \|\beta - \beta_0\| \leqslant Cn^{-1/2}\}$ 上进行的最小化，其中 $C > 0$ 为一常数。Härdle，Hall 和 Ichimura(1993)采用了这种方法。要求 β 位于一个集合有 $\beta - \beta_0 = O(n^{-1/2})$ 的假定可能显得过于强了；然而，给定 Ichimura(1993)的结果，我们知道 $\hat{\beta}$ 是 β 的一个$\sqrt{n}$一致估计量，我们可以看到 $S(\beta)$ 关于 β 的最小值有距 β_0 为 $O(n^{-1/2})$ 的假定。由这个假定以及对某些 $C_2 > C_1 > 0$，有 $h \in \mathcal{H}_n = \{h: C_1 n^{-1/5} \leqslant h \leqslant C_2 n^{-1/5}\}$，可以建立定理 8.1 的另一种证明。首先，可以证明非参数残差平方和可以写成(我们略去 $w_i = 1$ 和修剪示性函数来使符号简单化)

$$S_n(\beta) = \frac{1}{n}\sum_i [Y_i - \hat{G}_{-1}(X_i'\beta)]^2 = S(\beta) + T_n + o_p(1) \tag{8.9}$$

其中 $S(\beta) = n^{-1}\sum_i [Y_i - G(X_i'\beta)]^2$，其中

$$T_n = n^{-1}\sum_{i=1}^n [\hat{G}_{-i}(X_i'\beta_0) - g(X_i'\beta_0)]^2$$

是独立于 β 的一项且 $o_p(1)$ 表示在 $\beta \in \mathcal{B}_n$ 中一致等价于阶数为 $o_p(1)$ 的项。式(8.9)的证明可在 Härdle 等(1993)中找到，这事实上是考虑了一个更一般的设定，其中他们用交错鉴定法来同步选择 β 和 h。

257 在 8.12 节我们证明了 $S(\beta) = O(1)$。因此，对 β 来最小化式(8.9)渐近等价于最小化 $S(\beta)$。令 $\bar{\beta}$ 为最小化 $S(\beta)$ 的 β 值，那么通过 Taylor 展开，容易证明(见 8.12 节)$\bar{\beta}$ 满足下面一阶条件：

$$W_0(\bar{\beta} - \beta_0) = V_0 + (s.o.) \tag{8.10}$$

其中

$$W_0 = \sum_{i=1}^n u_i[g^{(1)}(X_i'\beta_0)]^2[X_i - E(X_i \mid X_i'\beta_0)][X_i - E(X_i \mid X_i'\beta_0)']$$

$$V_0 = \sum_{i=1}^n u_i f^{(1)}(X_i'\beta_0)[X_i - E(X_i \mid X_i'\beta_0)]$$

因此，通过一个标准的大数定律和中心极限定理，我们有

$$\sqrt{n}(\bar{\beta} - \beta_0) = (W_0/n)^{-1}n^{-1/2}V_0 + o_p(1) \xrightarrow{d} N(0, \Omega_0) \tag{8.11}$$

其中 Ω_l 和 Ω_o 相同，除了将其中的 $w(X_i)$ 替代为 1。

到目前为止，我们的正规化约束选择均为 $\|\beta\| = 1$。我们也可以根据需要来选择其他的正规化方法。不是假定 β 具有单位长度，我们可以假定 β 的首元是一连续变量，记为 $X_i = (X_{1i}, \tilde{X}_i')'$，$\tilde{X}_i$ 表示 X_i 去掉首元后剩下的部分。同样，定义 $\tilde{\beta}$ 为 $\tilde{X}_i$ 的系数，即 $\beta = (1, \tilde{\beta}')'$。这样我们可以选择 $\tilde{\beta}$ 来最小化

$$S_2(\tilde{\beta}) = n^{-1}\sum_i [Y_i - \hat{G}_{-i}(X_{1i} + \tilde{X}_i'\beta)]^2 w(X_i)\mathbf{1}(X_i \in A_x)$$

在此种情形下可以证明 $\sqrt{n}(\hat{\beta} - \tilde{\beta})$ 的渐近方差为 $\Omega_2 = V_2^{-1}\Sigma_2 V_2^{-1}$，其中

$$\Sigma_2 = E\{w(X_i)\sigma^2(X_i)(g_i^{(1)})^2$$
$$\times(\tilde{X}_i - E_A(\tilde{X}_i \mid X_i'\beta))(\tilde{X}_i - E_A(\tilde{X}_i \mid X_i'\beta))'\}$$

其中 $g_i^{(1)} = [\partial g(v)/\partial\tilde{\beta}]|_{v=X_i'\beta}$，$E_A(\tilde{X}_i|v) = E(\tilde{X}_i|X_A'\beta = v)$ 且在 $X \in A_x$ 条件下，X_A 具有 X 的分布，且 $V_2 = E[W(X_i)(g_i^{(1)})^2(\tilde{X}_i - E_A(\tilde{X}_i|X_i'\beta))(\tilde{X}_i - E_A(\tilde{X}_i|X_i'\beta))']$。

8.3　β的直接半参数估计量

258

Ichimura(1993)的半参数非线性最小二乘估计量在实际中计算可能会非常麻烦,这是由于目标函数可能是多峰的,于是就需要广泛搜索来排除局部极小值。本节讨论一个不需要迭代求解最优化问题的一个直接估计方法。这样计算的代价会比Ichimura提出的半参数非线性最小二乘方法要小;但是这并不是没有代价的,我们将会在本节最后讨论直接估计方法带来的一些问题。

8.3.1　平均导数估计量

当X_i为连续变量的向量时,我们也可以通过估计条件均值函数$E(Y|x)$的平均导数来估计单指标模型。由$E(Y|x)=g(x'\beta_0)$,我们得到

$$\frac{\partial E(Y\mid x)}{\partial x}=g^{(1)}(x'\beta_0)\beta_0 \tag{8.12}$$

因此,式(8.12)的平均值(平均导数)与β_0成比例,即

$$E\left[\frac{\partial E(Y\mid x)}{\partial x}\right]=E[g^{(1)}(x'\beta_0)]\beta_0\equiv C\beta_0 \tag{8.13}$$

也可以考虑一个加权平均导数形式

$$E\left[w(x)\frac{\partial E(Y\mid x)}{\partial x}\right]=E[w(x)g^{(1)}(x'\beta_0)]\beta_0\equiv C_2\beta_0 \tag{8.14}$$

它同样也与β_0成比例。

在核函数是可微的假定条件下,人们可以通过估计平均导数$E[\partial E(Y|x)/\partial x]$来估计$\beta_0$,如

$$\tilde{\beta}_{\text{ave}}=\frac{1}{n}\sum_{i=1}^{n}\frac{\partial\hat{E}(Y_i\mid X_i)}{\partial X_i} \tag{8.15}$$

其中下标ave表示这是平均导数估计量,$\hat{E}(Y_iX_i)=\sum_{j=1}^{n}Y_jK_{ij}/\sum_{j}K_{ij}$为$E(Y_i|X_i)$的局部常数核估计量,$K_{ij}=K((X_i-X_j)/a)$是一个乘积核函数,$a$是平滑参数向量。若选取$|\beta|=1$作为正规化标准,则尺度归一化可以通过$\tilde{\beta}_{\text{ave}}$除以 259 $|\tilde{\beta}_{\text{ave}}|=\{\sum_{j=1}^{q}\tilde{\beta}^2_{\text{ave},j}\}^{1/2}$来得到:若选取首个变量的系数为1作为正规化标准,则可以通过$\tilde{\beta}_{\text{ave}}$除以$\tilde{\beta}_{\text{ave},1}$来正规化。

对$\tilde{\beta}_{\text{ave}}$渐近分布求导所遇到的一个困难是随机分母的存在,这样就导致在某些点上分母可以任意接近于0。Rilstone(1991)提出用一个修剪函数来避免

"小分母问题",并建立式(8.15)中定义的 $\tilde{\beta}_{\text{ave}}$ 的 $\sqrt{n}$ 正态性结论。

当在 X 的支撑集边界上有 $f(x)=0$ 时(例如,如果 X 具有无界支撑),Powell,Stock 和 Stoker(1989)建议使用式(8.14)定义的加权平均导数估计量,其中权函数 $w(x)=f(x)$。那么,由分部积分可以得到

$$\begin{aligned}\delta &\stackrel{\text{def}}{=} E[f(X)]g^{(1)}(X)] \\ &= \int g^{(1)}(x)f^2(x)\,\mathrm{d}x \\ &= \int\left(\frac{\delta g(x'\beta_0)}{\delta x}\right)f^2(x)\,\mathrm{d}x \\ &= 0-2\int g(x'\beta_0)f(x)f^1(x)\,\mathrm{d}x \\ &= -2E[g(X'\beta_0)f^{(1)}(X)] \\ &= -2E[Yf^{(1)}(X)] \end{aligned} \tag{8.16}$$

可以用 $\hat{\delta}=-\dfrac{2}{n}\sum_i Y_i\hat{f}^{(1)}(X_i)$ 来估计式(8.16)。 (8.17)

其中 $\hat{f}^{(1)}(X_i)$ 为一 $q\times 1$ 向量,其第 s 项为 $\partial\hat{f}(X_i)/\partial X_{is}=n^{-1}\sum_{j=1}^{n}a_s^{-2}k^{(1)}((X_{is}-X_{js})/a_s)\prod_{t\neq s}^{q}a_t^{-1}k((X_{it}-X_{jt}/a_t)$。为了避免与半参数指标函数估计中的平滑参数 h 相混淆,这里我们用 a 来表示平均导数估计量中的平滑参数。注意通过选择 $w(x)=f(x)$,定义在式(8.17)中的 $\hat{\delta}$ 不再具有随机分母,于是也就不再需要引入修剪冗余参数。Härdle 和 Stoker(1989)考虑了基于未加权量 $E[g^{(1)}(X)]=-2E[Yf^{(1)}(X)/f(X)]$ 的平均导数估计量,他们使用了修剪方法来避免小随机分母的问题。Powell 等(1989)建立了定义在(8.17)式中的 $\hat{\delta}$ 的 $\sqrt{n}$ 正态性结果。

260

如正文所述,关键的假定是在 X 支撑集的边界上有 $f(x)=0$。其他的平滑条件和矩条件可以参考 Powell 等(1989),或者参考 Horowitz(1998,pp. 37—38)中的条件(a)—(d)。这些条件包括了未知函数 $g(\cdot)$ 与 $f(\cdot)$ 矩条件的平滑性,v 阶核函数的运用,以及当 $n\to\infty$ 时,$n\sum_{s=1}^{q}a_s^{2v}\to 0$ 与 $n(a_1\cdots a_q)\sum_{s=1}^{q}a_s^2\to\infty$。最后两个条件要求 $v>1+q/2$。由于 $q\geq 2$,这就要求使用高阶核函数($v>2$)。在一个单指标模型情形下,我们要求 $q\geq 2$,否则 β 是不可识别的。Powell 等(1999)证明了以下结果:

$$\sqrt{n}(\hat{\delta}-\delta)\to N(0,\Omega_{\text{PNS}}) \tag{8.18}$$

其中 $\Omega_{\text{PNS}}=4E[\sigma^2(X)f^{(1)}(X)f^{(1)}(X)']+4\text{var}(f(X)g^{(1)}(X))$。正规化的 β 向

量可由 $\hat{\delta}/|\hat{\delta}|$ 得到。

式(8.18)的证明需要用到可变核函数的 U 统计量分解。(一些相关内容见附录 A。)

在第2章我们讨论了使用局部多项式方法来估计一个未知条件期望函数和它的导数。由于在式(8.15)中给出的 $\tilde{\beta}_{\text{ave}}$ 是基于 $E(Y|x)$ 的一个局部常数核估计量的导数,我们也可以用局部多项式方法来估计 $\partial E(Y|x)/\partial x$。Li,Lu 和 Ullah(2003)考虑了这种方法,令 $\hat{g}^{(1)}(X_i)$ 表示 $g^{(1)}(X_i)$ 的核估计量,通过一个 m 阶局部多项式回归可得到它。Li,Lu 和 Ulah 建议使用

$$\hat{\beta}_{\text{ave}} = \frac{1}{n}\sum_i \hat{g}^{(1)}(X_i) \tag{8.19}$$

来估计 $\beta = E[g^{(1)}(X)]$。这里用 $\hat{\beta}_{\text{ave}}$ 直接估计 β,并没有假定在 X 的支撑集边界上有 $f(x)=0$。在这种意义上其与式(8.15)中定义的 $\tilde{\beta}$ 相似。然而不使用在其支撑边界上 $f(x)=0$ 的条件的代价是,正如 Li,Lu 和 Ullah 指出的,人们通常假定 X 的支撑集为一个紧集,且密度函数 $f(X)$ 是有界的并小于在 X 的支撑上的一个正的常数,即这些条件排除了无界支撑集的情形。在支撑集无界的情形下,需要引入一个修剪函数来剔除值很小的随机分母。在有界支撑和在其支撑上密度函数有界非零的假定下,就不再需要引入修剪函数。Li,Lu 和 Ullah 使用了 Masry(1996a)的一致收敛速度结果来处理随机分母问题。在 Powell 等(1989)中使用的类似的平滑条件和矩条件下,他们的有界假定被替换为紧支撑及密度函数在其支撑上是有界非零的假定,使用一个二阶核且 $n\to\infty$ 时,$n\sum_{s=1}^{q} a_s^{2m}\to 0, n(a_1\cdots a_q)\sum_{s=1}^{q} a_s^2/\ln(n)\to\infty$,其中 m 为局部多项式估计中多项式的阶数,q 为 x 的维数,Li,Lu 和 Ullah 建立了如下结论:

261

$$\sqrt{N}(\hat{\beta}_{\text{ave}} - \beta) \to N(0, \Phi + \text{var}(g^{(1)}(X))) \tag{8.20}$$

其中 $\Phi = E[\sigma^2(X)f^{(1)}(X)f^{(1)}(X)'/f^2(X)]$,$\beta = E[g^{(1)}(X)]$。

$\hat{\delta}$ 与 $\hat{\beta}_{\text{ave}}$ 方差相异是由于 $\hat{\delta}$ 估计的是 $\delta = E[f(X)g^{(1)}(X)]$,而 $\hat{\beta}_{\text{ave}}$ 估计的是 $\beta = E[g^{(1)}(X)]$。然而上面给出的 $\hat{\delta}$ 与 $\hat{\beta}_{\text{ave}}$ 均未正规化为有单位长度,如果应用正规化,由 $\hat{\delta}$ 与 $\hat{\beta}_{\text{ave}}$ 得到的正规化向量的方差将是相同的。这就是人们所期望的 Newey(1994b)给出的结论,他证明了半参数模型中 $\sqrt{n}$ 一致估计量的渐近方差与使用的特定的非参数估计方法是无关的,事实上,Newey 也证明了如果人们使用非参数级数方法而不是核方法时,平均异数估计量的渐近方差仍然是相同的。我们在第15章讨论级数方法。

式(8.20)的证明与式(8.18)的证明类似,都使用了 Powell 等(1989)发展

出的可变核的 U 统计量分解，也都广泛应用了 Masry(1996a)中的结果。这里我们不重复在 Li, Lu 和 Ullah(2003)包含的详细证明，我们只给出一个式(8.20)证明的简短提纲来给读者一个直观的理解。定义 $\bar{\beta}=n^{-1}\sum_{l=1}^{n}g^{(1)}(X_l)$，我们记

$$\sqrt{n}(\hat{\beta}_{\text{ave}}-\beta)=\sqrt{n}(\hat{\beta}_{\text{ave}}-\bar{\beta})+\sqrt{n}(\bar{\beta}-\beta)$$

262 注意由 Lindeberg-levy 中心极限定理有，$\sqrt{n}(\bar{\beta}-\beta)\xrightarrow{d}N(0,\text{var}(g^{(1)}(X)))$。可以证明 $\sqrt{n}(\hat{\beta}_a-\bar{\beta})\xrightarrow{d}N(0,\Phi)$ 依分布收敛。最后，以上两项是渐近独立的。因此，

$$\sqrt{n}(\hat{\beta}_a-\beta)\to N(0,\Phi+\text{var}(g^{(1)}(X)))\text{ 依分布收敛}$$

Hristache, Juditsky 和 Spokoiny(2001)提出一个迭代过程来改进原始的(非迭代)平均导数估计量。他们的想法是用向量 β 的先验信息来改进梯度估计的质量，通过在小的方向导数方向上扩展一个加权核，他们也证明了整个过程需要至多 $2\log(n)$ 次迭代。最终得到的估计量在相对适度的假定下是 $\sqrt{n}$ 一致的。

8.3.2 $g(\cdot)$的估计

我们用 β_n 来表示 β 或者 δ 一个一般的 $\sqrt{n}$ 一致估计量(即它可以为前面所定义的 $\tilde{\beta}$、$\hat{\beta}$、$\tilde{\beta}_{\text{ave}}$、$\hat{\beta}_{\text{ave}}$ 或者 $\hat{\delta}$)。有了 β_n，我们就能用

$$\hat{g}(x'\beta_n)=\frac{\sum_{j=1}^{n}Y_jK\left(\frac{(X_j-x)'\beta_n}{h}\right)}{\sum_{j=1}^{n}K\left(\frac{(X_j-x)'\beta_n}{h}\right)}\tag{8.21}$$

来估计 $E(Y|x)=g(x\beta_0)$。

由于 $\beta_n-\beta_0=O_p(n^{-1/2})$，其收敛于零的速度比标准非参数估计量要快，$\hat{g}(x'\beta_n)$ 的渐近分布与 β_n 被 β_0 替代的情形相同。因此，第2章定理2.2包括了这种情况，其中 $q=1$(由于 $v=x'\beta_0$ 为一标量)。因此，我们有

推论 8.1 假定 $\beta_n-\beta_0=O_p(n^{-1/2})$，在与定理2.2给出的类似的条件下，我们有

$$\sqrt{nh}[\hat{g}(x'\beta_n)-g(x'\beta_0)-h^2B(x'\beta_0)]\xrightarrow{d}N(0,\kappa\sigma^2(x'\beta_0)/f(x'\beta_0)),\tag{8.22}$$

其中 $B(X'\beta_0)$ 定义来自式(2.8)。

上面讨论的直接平均导数估计方法只适用于 x 为 q 维连续变量向量情形，263 这是由于对离散变量取导数没有意义。Horowitz 和 Härdle(1996)讨论直接(非

迭代)估计怎样可以扩展到在 x 的某些元素为离散时的情形。Horowitz(1998)给出了此方法的一个很好的综述。有关详细讨论我们推荐读者参考 Horowitz 和 Härdle(1993)以及 Horowitz(1998,pp. 41—48)。

直接平均导数估计方法的优点是人们可以不使用非线性迭代过程而直接来估计 β_0 与 $g(x'\beta_0)$。这样带来的计算简便化在大样本情形下非常有吸引力;然而,随之产生了一个潜在的有限样本问题。直接估计量在初始估计阶段都涉及了多维非参数估计,而我们知道非参数估计量存在维数诅咒的问题。在第二阶段,多维非参数估计量对所有样本点进行平均,得到一个 β_0 的$\sqrt{n}$一致估计量。由于第二阶段的估计具有参数的$\sqrt{n}$速度的收敛性,且 x 的维数不影响第二阶段得到的平均导数估计量的收敛速度,因此,渐近来说,维数诅咒问题不存在了。但是,在有限样本的实际应用中,除非样本容量非常大,否则在第一阶段不准确的估计可能会影响第二阶段估计量的准确性。所以,在样本容量相当大时,直接估计由于其计算上的简便更具有吸引力。然而在小样本情形下,Ichimura(1993)的迭代方法可能更具有吸引力,这是因为其避开了高维非参数估计的使用。

Carroll,Fan,Gijbels 和 Wand 和(1997)提出了一个与第 7 章内容(即部分线性模型)和本章内容(即单指标模型)都相关联的方法。特别地,Carroll 等考虑了一个广义的部分线性单指标模型的问题,这种模型作为一种特殊情形既包含了部分线性模型也包含了单指标模型。

8.4　窗宽选择

8.4.1　Ichimura 方法的窗宽选择

Ichimura(1993)方法通过(迭代)非线性最小二乘方法估计 β,其中未知函数 $g(X_i'\beta_0)$用式(8.7)中定义的非参数核估计量 $\hat{g}(X_i'\beta)$来代替。选择的平滑参数要满足条件 $nh^8\to0$,且当 $n\to\infty$ 时有 $\ln(h)/[nh^{3+3/(v-1)}]\to0$,这里 $v\geqslant3$ 是一正整数,其具体值依赖于 Y 的一定数目有限矩的存在以及未知函数 $g(\cdot)$ 的光滑性。可允许的平滑参数取值范围考虑到了最优平滑,即 $h=O(n^{-1/5})$。因此,Härdle 等(1993)建议用 Ichimura 的非线性最小二乘交错鉴定法来同步选取 h 和 β。特别地,他们建议最小化 264

$$M(\beta,h)=\sum_i[Y_i-\hat{G}_{-i}(X_i'\beta,h)]^2\mathbf{1}(X_i\in A_\delta)\qquad(8.23)$$

来同步选取 h 和 β。其中 $\hat{G}_{-1}(X_i'\beta,h)=\hat{G}_{-1}(X_i'\beta)$如式(8.7)所定义,而 A_δ 即前面所介绍的修剪集。

在一些正则条件下，包括 Y 具有任意阶的有限矩，使用二阶核函数，未知函数 $g(x'\beta)$ 和 $p(x'\beta)$ 是二次可微的，$f(x)$ 在 A_δ 上是非零有界的，以及假定 $\beta\in\mathcal{B}_n\{\beta:|\beta-\beta_0|\leqslant C_0n^{-1/2}\}$，其中 $h\in H_n=[C_1n^{-1/5},C_2n^{-1/5}]$，$C_0,C_2>C_1$ 是三个正的常数，然后，Härdle 等（1993）证明了 $M(\beta,h)$ 能分解为如下形式

$$M(\beta,h)=M(\beta)+T(h)+\{\text{比 } T(h) \text{ 有更小的阶且不依赖于 } \beta \text{ 的项}\}$$
$$+\{\text{比 } M(\beta) \text{ 和 } T(h) \text{ 有更小的阶的项}\} \tag{8.24}$$

其中

$$M(\beta)=\sum_i[Y_i-g(X_i'\beta)]^2 1(X_i\in A_x)$$

$$T(h)=\sum_i[\hat{G}_{-i}(X_i'\beta_0)-g(X_i'\beta_0)]^2$$

且 $\hat{G}_{-i}(X'\beta_0)$ 如式（8.7）所定义但是将 β 替换为 β_0。

因此，对 $(\beta,h)\in\mathcal{B}_n\times H_n$ 同步最小化 $M(\beta,h)$ 等价于先对 $\beta\in\mathcal{B}_n$ 最小化 $M(\beta)$ 再对 $h\in H_n$ 来最小化 $T(h)$。

令 $(\hat{\beta},\hat{h})$ 表示最小化式（8.23）得到的估计量。$\hat{\beta}$ 的渐近分布由定理 8.1 给出，对于 $\hat{h}$，由于使用的是二阶核函数，用第 2 章讨论的非参数估计量的标准均方误差计算容易证明非随机目标函数 $E[T(h)]$ 的首项等于 $A_1h^4+A_2(nh)^{-1}$，其中 A_1 与 A_2 是两个正的常数，因此，$h_0=[A/(4A_2)]^{1/5}n^{-1/5}=O(n^{-1/5})$ 最小化了 $A_1h^4+A_2(nh)^{-1}$。Härdle 等（1993）证明了 $\hat{h}/h_0$ 依概率收敛于 1。

265

我们现在简要比较一下 Ichiruma（1993）与 Härdle 等（1993）中使用的正则条件。用于证明定理 8.1 的正则条件与建立式（8.24）的正则条件是有所不同的。例如，在定理 8.1 中使用的条件要求一个更高阶的核，而 Härdle 等使用一个具有最优平滑参数 $h=O(n^{-1/5})$ 的二阶核来推导式（8.11）。在式（8.24）中，最小化是在限制收缩集 $(\beta,h)\in\mathcal{B}_n\times H_n$ 中进行的，故有 $\beta-\beta_0=O(n^{-1/2})$。这个条件使得核估计的偏差比其在定理 8.1 的条件下要小，因此人们可以得出式（8.11）而不必再求助于高阶核。正则条件间的另一个差异是使用了一个更强的矩条件，即 Y 有任意阶的矩，来得到式（8.11）。这是因为 Härdle 等需要式（8.11）中的低阶项 在 $\mathcal{B}_n\times H_n$ 上保持统一。证明一致的收敛速度通常需要较强的矩条件，然而在定理 8.1 中，h 被当作非随机的，最小化是只针对 β 进行的，故一个较弱的矩条件即可满足要求。

8.4.2 直接估计方法中窗宽的选择

对于直接平均导数估计方法，β_0 的估计涉及一阶导数的 q 维多元非参数估计。Härdle 和 Tsybakov（1993）提出选择平滑参数 $a_1,\cdots,a_q$ 来最小化 $\hat{\delta}$ 的均方误差，也即选择 h 来最小化 $E[|\hat{\delta}-\delta|^2]$。Härdle 和 Tsybakov 证明了渐近最优

窗宽具有如下形式(对所有 $s=1,\cdots,q$):

$$a_s = c_s n^{-2/(2q+v+2)}$$

其中 c_s 是一常数,v 为核的阶数,q 为 x 的维数。

Powell 和 Stoker(1996)给出了一个估计 c_s 的方法,而 Horowitz(1998, pp. 50—52)考虑了基于自举法抽样来选择 a_s。

在已经选择最优的 a_s 后就可以得到 β 的一个平均导数估计量。用 β_n 来表示一个一般的估计量。然后,人们通过在式(8.7)所定义的 $\hat{g}(x'\beta_n,h)\equiv\hat{g}(x'\beta_n)$ 来估计 $E[Y|x]=g(x'\beta_0)$。与标量指标 $x'\beta_n$ 相关联的平滑参数 h 可以通过最小二乘交错鉴定法来选择,即选择 h 来最小化 $\sum_{i=1}^{n}[Y_i-\hat{g}_{-1}(X'\beta_n,h)]^2$。在某些正则条件下,交错鉴定法选择的 h 是$O_p(n^{-1/5})$阶的。 266

人们也可以把部分线性模型和单指标模型结合在一起得到一个"部分线性单指标模型",具有形式 $E(Y|X,Z)=X'\alpha+g(Z'\beta)$。对部分线性单指标模型的估计见 Carroll 等(1997),Xia,Tong 和 Li(1999)以及 Liang 和 Wang(2005)。

8.5　Klein 和 Spady 的估计量

当单指标模型是从二元选择模型(8.2)导出时,并且在 ε_i 与 X_i 独立的假定下,Klein 和 Spady(1993)提出了用最大似然方法来估计 β。估计的对数似然函数为

$$\mathcal{L}(\beta) = \sum_i (1-Y_i)\ln(1-\hat{g}(X_i'\beta)) + \sum_i Y_i \ln[\hat{g}(X_i'\beta)] \quad (8.25)$$

其中 $\hat{g}(X_i'\beta)$定义见式(8.7)。Klein 和 Spady 提出,关于 β 最大化式(8.25)得到 β 的半参数最大似然估计量,记为 $\hat{\beta}_{KS}$。类似于 Ichiruma(1993)的估计量,最大化的实现必须通过数值求解由式(8.25)得到的一阶条件。

在一些正则条件下,包括引入一个修剪函数来剔除 X_i 支撑集边界附近的观测值,使用高阶核,Klein 和 Spady(1993)证明了 $\hat{\beta}_{KS}$是$\sqrt{n}$一致的且具有渐近正态分布如下

$$\sqrt{n}(\hat{\beta}_{KS}-\beta) \to N(0,\Omega_{KS})$$

其中

$$\Omega_{KS} = \left[E\left\{\frac{\partial P}{\partial\beta}\left(\frac{\partial P}{\partial\beta}\right)'\left[\frac{1}{P(1-P)}\right]\right\}\right]^{-1}$$

$P=P(x'\beta>-\varepsilon)=F_{\varepsilon|x}(x'\beta)$,其中 $F_{\varepsilon|x}(\cdot)$是 ε_i 在 $X_i=x$ 条件下的累积分布函数。

267 Klein 和 Spady(1993)也证明了他们提出的估计量在其渐近方差达到半参数有效边界的意义上是半参数有效的。

我们将半参数估计量 Ω_{KS} 的渐近方差与对应参数估计量 Ω_{nls} 的渐近方差进行比较。参数模型具有两个额外的参数，$\eta=(\gamma_0,\gamma_1)'$。将参数 γ 分拆为 $\gamma=(\eta'\beta')'$，那么 $n^{1/2}(\hat{\beta}_{nls}-\beta)$ 的渐近方差为 $V_{nls}\stackrel{\text{def}}{=}(\mathcal{I}^0_{\beta\beta}-\mathcal{I}^0_{\beta\eta}(\mathcal{I}^0)^{-1}\mathcal{I}^0_{\eta\beta})^{-1}$。将其与 V_{KS}^{-1} 对比，可以证明（例如，Pagan 和 Ullah（1999, p. 278））若 $E(X_i|X'\beta)=c_0+c_1(X_i'\beta)$，这里 c_0 和 c_1 为两个 $q\times1$ 的常数向量，那么有 $V_{KS}^{-1}=V_{nls}^{-1}$，或者等价地有 $V_{KS}=V_{nls}$。

也就是说，半参数估计量是渐近地与参数非线性最小二乘估计量一样有效的，这里参数非线性最小二乘估计量是基于当 $E(X_i|X_i'\beta)$ 对 $X_i'\beta$ 具有线性形式（“一阶有效性”）时 $g(\cdot)$ 的真实函数形式是已知的条件得出的。这与部分线性模型的情形是相似的。然而，当 $E(X_i|X_i'\beta)$ 不是 $X_i'\beta$ 的一个线性函数时，可以证明 $V_{KS}-V_{nls}$ 是正定的。因此半参数估计量与基于 $g(\cdot)$ 真实函数形式的参数非线性最小二乘估计量相比是渐近弱有效的，渐近方差 $V_{KS}-V_{nls}$ 是正定的，除非 $E(X_i|X_i'\beta)=c_0+c_1(X_i'\beta)$。此外，由于 V_{KS} 已经达到半参数有效边界，此种情形下结果不能被改进。与参数非线性最小二乘估计量相比，半参数模型的有效性缺失是由于 $g(\cdot)$（或者等价地，$F_{\varepsilon|x}(\cdot)$）的函数形式未知。当然，实际中 $g(\cdot)$ 的真实函数形式普遍是未知的，这样半参数估计量对于 $g(\cdot)$ 函数形式被误设是稳健的。

8.6 Lewbel 的估计量

Lewbel(2000)考虑了如下二元选择模型：

$$Y_i=\mathbf{1}(v_i+X_i'\beta+\varepsilon_i>0) \tag{8.26}$$

其中 v_i 是一个（特殊）连续回归元，其系数被正规化为 1，X_i 的维数为 q。令 $f(v|x)$ 表示给定 X_i 时 v_i 的条件密度函数，$F_\varepsilon(\varepsilon|v,x)$ 表示给定 (v_i,X_i) 时 ε_i 的条件累积分布函数。在 $F_e(\varepsilon|v,x)=F_e(x)$ 的条件下，即以 x 为条件，ε 与特殊回归元 V_i 是相互独立的，且 $E[X_i\varepsilon_i]=0$，Lewbel 证明 268

$$\beta=[E(X_iX_i')]^{-1}E[X_i\tilde{Y}_i] \tag{8.27}$$

式(8.26)建议人们通过 $\tilde{Y}_i$ 对 X_i 回归来估计 β。$\tilde{Y}_i$ 涉及了未知量 $f(v_i|X_i)$，可以通过第 5 章介绍的非参数核方法来得出 $f(v_i|X_i)$ 的一致估计。令 $\hat{\beta}$ 表示相应的 β 的可行估计量，Lewbel(2000)建立了其提出的 β 的估计量的 $\sqrt{n}$ 正态性结果。

Lewbel(2000)在 ε_i 与 X_i 相关,即 $E(\varepsilon_i X_i \neq 0)$ 的情形下,进一步扩展了其所得到的结果。假定存在一个 p 维工具变量向量 Z_i 满足 $E(\varepsilon_i Z_i)=0$,$E(Z_i X'_i)$ 非奇异,且 $F_{\varepsilon x}(\varepsilon, x \mid v, z) = F_{\varepsilon x}(\varepsilon, x \mid z)$,其中 $F_{\varepsilon x}(\varepsilon, x \mid \cdot)$ 表示(ε, x)以数据 · 为条件的分布函数,Lewbel 证明

$$\beta = [E(Z_i X_i')]^{-1} E\left[Z_i \frac{Y_i - \mathbf{1}(v_i > 0)}{f(v_i \mid Z_i)}\right]$$

因此,人们可以用样本均值替换上式中的期望并用 $f(v_i \mid Z_i)$ 的一个一致估计量替换 $f(v_i \mid Z_i)$ 来估计 β。

上述方法可以扩展到定义如下的有序响应模型

$$Y_i = \sum_{j=0}^{J-1} j\mathbf{1}(a_j < v_i + X_i'\beta + \varepsilon_i \leqslant a_j + 1) \tag{8.28}$$

其中 $a_0 = -\infty$, $a_J = +\infty$。响应变量 Y_j 在 $\{0,1,\cdots,J-1\}$ 中取值,且当 $v_i + X_i\beta + \varepsilon_i$ 位于 a_j 与 a_{j+1} 之间时,有 $Y_i = j$。令 $X_{1i} = 1$(截距项),同时不失一般性令 $\beta_i = 0$(否则可以重新定义 a_j 为 $a_j - \beta_1$)。令 $Y_{ji} = \mathbf{1}(Y_i \geqslant j)$,$j = 1,\cdots,J-1$,定义 $\Delta = [E(X_i X_i')]^{-1}$,同时令 Δ_j 为 Δ 的第 j 行。Lewbel(2000)证明

$$\alpha_j = -\Delta_1 E\left(X_i \frac{Y_{ji} - \mathbf{1}(v_i > 0)}{f(v_i \mid X_i)}\right), \quad j = 1,\cdots,J-1 \tag{8.29}$$

$$\beta_l = -\Delta_j E\left(X_i \frac{\sum_{j=1}^{J-1} Y_{ji}/(J-1) - \mathbf{1}(v_i > 0)}{f(v_i \mid X_i)}\right), \quad l = 2,\cdots,q \tag{8.30}$$

从式(8.29)与式(8.30)中可以容易得出 α_j 与 β_l 的可行估计量,即用样本均值替换上式中的期望并用 $f(v_i \mid X_i)$ 的一个一致估计量来替换 $f(v_i \mid X_i)$。Lewbel(2000)建立了所得估计量的渐近正态性结果。Lewbel 进一步证明了其结果能够扩展来处理多项选择、部分线性潜变量模型以及阈值和审查回归模型。

269

8.7 Manski 的最大得分估计量

Manski(1975)的最大得分[①]估计量涉及选择 β 来最大化以下目标函数:

$$S_M(\beta) = \sum_{i=1}^{n} Y_i \mathbf{1}(X_i'\beta \geqslant 0) + (1 - Y_i)\mathbf{1}(X_i'\beta < 0) \tag{8.31}$$

这个估计量寻求的是最大化正确预测的数目。对 $Y_i = 1$,若 $X_i'\beta \geqslant 0$,有 $S_M = 1$,若 $X_i'\beta < 0$,则有 $S_M(\beta) = 0$。正确的预测获得权数 1,一个不正确的预测获

① 注意 Manski 这里使用的术语"得分"(score)类似于"记分",如棒球比赛中得分的意思,而不是指统计上的 score 函数(如对数似然函数梯度之和)。

得权数 0。类似地，对于 $Y_i=0$，若 $X_i'\beta<0$，有 $S_M=0$，若 $X_i'\beta\geqslant 0$，则有 $S_M=-1$。这种情形下，正确的预测获得权数 1，不正确的预测获得的权数为 -1。Manski(1975)在 $\text{median}(Y_i|X_i)=X_i'\beta$(或者 $\text{median}(\varepsilon_i|X_i)=0$)，$\beta$ 的首元为 1，x 的首元为连续变量的假定下，证明了 $\hat{\beta}$ 的强一致性。Kim 和 Pollard(1990)证明了最大得分估计量具有收敛速度 $n^{-1/3}$，而不是通常的 $n^{-1/2}$。由于目标函数是非连续的，因此渐近理论中的标准 Taylor 级数展开方法不能应用在最大得分估计量上。Kim 和 Pollard 证明了 $n^{-1/3}(\hat{\beta}_{m\text{-score}}-\beta)$ 的极限分布是一个带有二次漂移的多维布朗运动最大值。这个渐近分布非常复杂因此不便应用于推断。Manski 和 Thompson(1986)提出了使用自举法来逼近 $\hat{\beta}_{m\text{-score}}$ 的渐近分布。自举法实施起来比较简便，Manski 和 Thompson 给出的模拟证明他们提出的自举法在有限样本应用中表现良好。

270

8.8 Horowitz 的平滑最大得分估计量

尽管 Manski(1975)证明了其给出的最大得分估计量在比较弱的分布假定下是一致的，但它的收敛速度较慢且渐近分布比较复杂。Horowitz(1992)提出了一个修正的最大得分估计量，可以通过最大化 Manski 的得分函数的平滑形式来得到，估计量逼近$\sqrt{n}$的速率依赖于某些平滑性假定的强度。实质上，Manski 方法的问题主要在于在式(8.31)中使用的示性函数缺少连续性。Horowitz 提出用一个保留了示性函数本质特征的二次连续可微函数来代替示性函数$\mathbf{1}(A)$。Horowitz 通过最大化如下平滑目标函数来估计 β：

$$\max_{\beta\in B,\beta_1=1} S_{sms}(\beta) \equiv \frac{1}{n}\sum_{i=1}^{n}(2Y_i-1)G\left(\frac{X_i'\beta}{h}\right) \tag{8.32}$$

其中 $G(\cdot)$ 是 p 次连续可微的累积分布函数 $h=h_n>0$，且当 $n\to\infty$ 时有$h\to 0$。易知当 $h\to 0$ 时，$G(X_i'\beta/h)\to(X_i'\beta\geqslant 0)$。例如，人们可以选择 $G(x)=\int_{-\infty}^{x}k(v)\,\mathrm{d}v$，$k(\cdot)$ 为一个 $p-1$ 次可微核函数。在一些正则假定下，包括给定 $x'\beta$ 关于 y 的条件密度函数的一些平滑条件，Horowitz 证明了其所提出的平滑最大得分估计量，记为 $\hat{\beta}_{sm\text{-score}}$，具有与 $\hat{\beta}_{sm\text{-score}}-\beta=(nh)^{-1/2}$相同的收敛速度且服从渐近正态分布。对某些 $0<c<\infty$，若 $h=(c/n)^{1/(2p+1)}$，那么 $\hat{\beta}_{sm\text{-score}}-\beta$ 具有和 $n^{-p/(2p+1)}$ 相同的收敛速度，如果 p 足够大，收敛速度将会非常接近 $n^{-1/2}$。

8.9 Han 的最大秩估计量

不同于式(8.31)给出的最大化得分函数,Han(1987)考虑了最大化二元结果 Y_i 与指标函数 $X_i'\beta$ 之间的秩相关性。这是通过对下式进行最大化得到的:

$$G_H(\beta) = \frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}^{n}\mathbf{1}(Y_i \geqslant Y_j)\mathbf{1}(X_i'\beta \geqslant X_j'\beta), \qquad (8.33)$$

271

这里的求和是在离散元素 $\{i,j\}$ 所有 $\begin{bmatrix} n \\ 2 \end{bmatrix} = n(n-1)/2$ 个组合上进行的。一个简单的原理促发了这个估计量。F 的单调性以及 ε_i 与 X_i 之间的独立性保证了

$$P(Y_i > Y_j \mid X_i, X_j) \geqslant P(Y_i \leqslant Y_j \mid X_i, X_j), X_i'\beta_0 \geqslant X_j'\beta_0$$

也就是说,当 $X_i'\beta_0 \geqslant X_j'\beta_0$ 时,$Y_i > Y_j$ 的可能性更大,Han(1987)证明了 $\beta = \beta_0$ 最大化了 $E[G_H(\beta)]$,β_0 为 β 的真实值 。Han 进一步建立了其提出的最大秩相关(MRC)估计量的强一致性,但是他并没有提供其给出的最大秩相关估计量的渐近分布。给出最大秩相关估计量极限分布的主要困难来自目标函数 $G_H(\beta)$ 的非平滑性。注意 $G_H(\beta)$ 是一个二阶 U 统计量(或者称为 U 过程,指标为 β)。使用 U 统计量分解以及退化 U 过程的一致界限,Sherman(1993)证明了最大秩相关估计量是 $\sqrt{n}$—致的且具有渐近正态分布。

到目前为止,我们考虑的都是半参数二元选择模型,模型中误差项的分布是非参数模式的,线性指标 $x'\beta$ 是模型的参数部分。Matzkin(1992)考虑了一个更一般的二元选择模型,模型中外生变量的系统函数与随机误差项分布均不使用任何参数结构。Matzkin 给出了识别条件且证明了她的非参数最大似然估计量的一致性。

8.10 多项式离散选择模型

Pagan 和 Ullah(1999,pp. 296—299)给出了一个关于多项式离散选择模型半参数估计很好的综述。对于下述讨论,考虑一个个体面临 $J(J>2)$ 种选择的情形。定义 $Y_{ij}=1$ 为个体 i 选择了第 j 种选择($j=1,\cdots,J$),否则 $Y_{ij}=0$。令 $F_{ij}=P(Y_{ij}=1 \mid X_i) \equiv E(Y_{ij} \mid X_i)$;那么多种选择方程为

$$Y_{ij} = F_{ij} + \varepsilon_{ij} \qquad (8.34)$$

似然函数为

272

$$\sum_{i=1}^{n}\sum_{j=1}^{J} Y_{ij}\ln F_{ij} \qquad (8.35)$$

参数方法指定了 F_{ij} 的函数形式。例如，设定下式可得到一个多项式 Logit 模型：

$$F_{ij} = \exp(X'_{ij}\beta) / \sum_{j=1}^{J} \exp(X'_{ij}\beta)$$

半参数方法设定 $F_{ij} = E(Y_{ij} | X_i) = E(Y_{ij} | v_{i1}, \cdots, v_{iJ}) = g(v_{i1}, \cdots, v_{iJ})$，其中 $g(\cdot)$ 的函数形式是未知的，且 $v_{ij} = X'_{ij}\beta_j$。估计过程与前面讨论过的半参数单指标模型类似。Ichimura 和 Lee(1991)将 Ichimura(1993)的方法拓展到多指标情形，并且导出了相应半参数最小二乘估计量的渐近分布。Lee(1995)提出使用半参数最大似然方法来估计多指标模型(8.35)。Ai(1997)考虑了一般化的半参数最大似然方法，这种方法可以应用于多种半参数模型，如作为特例的多指标模型以及部分线性模型。我们接下来讨论 Ai 的一般化方法。

8.11 Ai 的半参数最大似然方法

Ai(1997)考虑了一个一般化的半参数最大似然估计方法。令 $q(Y|X, \theta_0, g_0)$ 为给定 X 下 Y 的条件概率密度函数，其中 θ_0 为一有限维的参数(参数部分)，$g_0(\cdot)$ 为无限维的未知函数。Ai 进一步假定条件密度满足一个指标约束，即存在一些已知函数 $v_i(z, \theta)$ 与 $v_2(z, \theta)$，使得

$$q(Y | X, \theta_0, g_0) = J(Z, \theta_0) f(v_1(Z, \theta_0) | v_2(X, \theta_2), \theta_0) \quad (8.36)$$

其中 $f(\cdot | \cdot, \theta)$ 是对任意 θ 给定 v_2 时 v_1 的条件密度，$J(z, \theta)$ 为一已知的从 $v_1(z, \theta)$ 到 y 的雅可比转换矩阵。

273

模型(8.36)包含了许多人所熟知的半参数模型作为特例。例如，考虑一个部分线性回归模型 $\eta_1(Y) = X'_1\theta_1 + \eta_2(X'_2\theta_2 + X_3) + u$，其中 $\eta_1(\cdot)$ 与 $\eta_2(\cdot)$ 的函数形式是未知的，u 独立于 $x = (X_1, X_2, X_3)$ 且其密度函数 $\eta_3(\cdot)$ 是未知的。$\theta = (\theta_1, \theta_2)$ 是模型的参数部分，$\eta = (\eta_1, \eta_2, \eta_3)$ 是模型的非参数部分。例子中给定 x 下 y 的条件密度数为

$$q(Y | X, \theta, \eta) = \eta_3[\eta_1(Y) - X'_1\theta_1 - \eta_2(X'_2\theta_2 + X_3)] | \eta_1^{(1)}(y) | \quad (8.37)$$

另外它也是给定 $v_2 = (X'_1\theta_1, X'_2\theta_2 + X_3)$ 下 $v_1 = Y$ 的条件密度函数，其中 $\eta_1^{(1)}(y)$ 表示 $\eta_1(y)$ 对 y 的导数，$|\eta_1^{(1)}(y)|$ 为雅可比矩阵。

对于 Ichimura 和 Lee(1991)中的部分线性多指标模型，$\eta_1(Y) = Y$(于是雅可比矩阵 $J = 1$)，$v_1 = Y - X'_1\theta_1$，$v_2 = X'_2\theta$，此时式(8.37)变为(其中雅可比矩阵 = 1)

$$f(v_1 | v_2) = \eta_3(Y - X'_1\theta - \eta_2(X'_2\theta_2)) \quad (8.38)$$

其中 $\eta_3(\cdot)$ 为 $u = Y - X_1'\theta_1 - \eta_2(X_2'\theta)$ 的密度函数，$\eta_2(X_2'\theta_2) = E[Y - X_1'\theta_1 | X_1'\theta]$。若人们使用在式(8.38)中的 $\theta_1 = 0$，模型退化为 Ichimura(1993)给出的单指标模型(见第 8 章)。

对于第 7 章中讨论的部分线性模型，$\eta_1(Y) = Y, v_1 = Y = X_1'\theta_1, \theta_2 = 0, v_2 = X_3$，那么式(8.37)变为($J = 1$)

$$f(v_1 \mid v_2) \equiv \eta_3(Y - X_1'\theta - \eta_2(Y_3))$$

其中 $\eta_3(\cdot)$ 为 $u = Y - X_1'\theta_1 - \eta_2(X_3)$ 的密度函数，$\eta_2(X_3) = E[Y - X_1'\theta | X_3]$。

我们现在讨论如何估计这一族模型。定义 $m(x, \theta, f) = \partial\ln[q(y|x,\theta,f)]/\partial\theta$。若 f 的函数形式是已知的，我们可以通过对 θ 解如下得分方程来估计 θ:

$$S_n(\tilde{\theta}) = \sum_{i=1}^{n} m(Z_i, \tilde{\theta}, f) = 0 \tag{8.39}$$

由标准大数定律和中心极限定理可以得出 $\tilde{\theta}$ 的$\sqrt{n}$正态性。当 f 未知时，Ai(1997)建议用非参数核方法来估计 f。令 $f_1(v(z,\theta),\theta)$ 为 v 的联合密度函数，$f_2(v_2(x,\theta),\theta)$ 为 v_2 的边际密度函数，我们有

$$f(v_1(z,\theta) \mid v_2(x,\theta),\theta) = \frac{f_1(v(z,\theta),\theta)}{f_2(v_2(x,\theta),\theta)}$$

将其代入 $m(z,\theta,f) = \partial\ln[q(y|x,\theta,f)]/\partial\theta$ 有

274

$$m(z,\theta,f) = m_1(z,\theta,f_1,f_2)m_2(z,\theta,f_1,f_2)$$

其中

$$\begin{aligned} m_1(z,\theta,f_1,f_2) = & \frac{\mathrm{d}\ln|J(z,\theta)|}{\mathrm{d}\theta} f_1[v(z,\theta),\theta] f_2[v_2(z,\theta),\theta] \\ & + \frac{\mathrm{d}f_1[v(z,\theta),\theta]}{\mathrm{d}\theta} f_2[v_2(z,\theta),\theta] \\ & - \frac{\mathrm{d}f_2[v_2(z,\theta),\theta]}{\mathrm{d}\theta} f_1[v(z,\theta),\theta] \end{aligned}$$

$$m_2(z,\theta,f_1,f_2) = \frac{1}{f_1[v(z,\theta),\theta] f_2[v_2(z,\theta),\theta]}$$

令 q 与 q_2 分别表示 v 和 v_2 的维数，那么 $f_i(\theta) \equiv f(v_1(Z_i,\theta) | v_2(X_i,\theta),\theta)$ 的核估计量由下式给出：

$$\hat{f}(v_1(Z_i,\theta) \mid v_2(X_i,\theta),\theta) = \frac{\hat{f}_1(v,\theta)}{\hat{f}_2(v_2,\theta)} \tag{8.40}$$

其中

$$\hat{f}_1(v,\theta) = \frac{1}{n}\sum_{j=1}^{n} K_{h_1}\left(\frac{v - v(Z_j,\theta)}{h_1}\right) \tag{8.41}$$

$$\hat{f}_2(v_2,\theta) = \frac{1}{n}\sum_{j=1}^{n} K_{h_2}\left(\frac{v_2 - v_2(X_j,\theta)}{h_2}\right) \tag{8.42}$$

其中乘积核定义为

$$K_h\left(\frac{v - v(Z_j,\theta)}{h_1}\right) = \prod_{s=1}^{q} h_s^{-1} k\left(\frac{v_s - v_s(Z_j,\theta)}{h_s}\right)$$

$$K_{h_2}\left(\frac{v_2 - v_2(X_j,\theta)}{h_2}\right) = \prod_{s=1}^{q_2} h_{2s}^{-1} k\left(\frac{v_{2s} - v_{2s}(X_j,\theta)}{h_{2s}}\right)$$

那么可得到$\hat{f}_i(\theta) = f_i(\theta) \equiv f(v_1(Z_j,\theta) | v_2(X_i,\theta),\theta)$的一个估计量为

$$\hat{f}_i(\theta) = \frac{\hat{f}(v(Z_j,\theta),\theta)}{\hat{f}_2(v_2(X_i,\theta),\theta)}$$

275 最后，解如下一阶条件可以得到 θ 的估计量 $\hat{\theta}$:

$$S_n(\tilde{\theta}) = \sum_{i=1}^{n} m(Z_i,\tilde{\theta},\hat{f}_i(\theta)) = 0 \tag{8.43}$$

对于单指标模型的例子，有雅可比矩阵 $=1$，$\theta_{10}=0$，$v_1 = y$，X_3 为空。我们记 $X_2 = x$，$\theta_2 = \theta$，那么有 $v_2 = x'\theta$。此时我们有

$$f_{1i} = f_1(v_i,\theta) \equiv f_1(Y_i, X_i'\theta) = n^{-1}\sum_{j=1}^{n} \mathbf{1}(Y_j = Y_i) K_h\left(\frac{(X_j - X_i)'\theta}{h}\right)$$

$$f_{2i} = f_2(v_{2i},\theta) \equiv f_2(x'\theta) = n^{-1}\sum_{j=1}^{n} K_h\left(\frac{(X_j - X_i)'\theta}{h}\right)$$

将这些代入

$$m(Z_i,\theta,\hat{f}_{1i},\hat{f}_{2i}) = m_1(Z_i,\theta,\hat{f}_{1i},\hat{f}_{2i}) m_2(Z_i,\theta,\hat{f}_{1i},\hat{f}_{2i})$$

我们就得到与 Klein 和 Spady(1993)的估计量一模一样的一阶条件[①]；我们把对此结论的验证作为习题(见习题 8.4)。

Ai(1997)证明了可行估计量 $\hat{\theta}$ 与非可行估计量 $\tilde{\theta}$ 具有相同的渐近分布，且对于此类半参数模型 $\hat{\theta}$ 在其渐近方差的逆等于半参数有效边界的意义上是半参数有效的。

8.12 定理 8.1 的证明概要

我们首先推导出式(8.45)。

注意有 $G(X_i'\beta) = E[g(X_i'\beta_0) | X_i'\beta]$，对 $g(X'\beta_0)$ 在 $\beta_0 = \beta$ 处进行 Taylor 展

① 为了表达的简单，我们忽略了修剪函数；对引入修剪函数的详细讨论见 Ai(1997, p.938)。

开，由 $\beta_0-\beta=O(n^{-1/2})$，我们有

$$\begin{aligned} g(X_i'\beta_0)-G(X_i'\beta) &\equiv g(X_i'\beta_0)-E[g(X_i'\beta_0)\mid X_i'\beta] \\ &= g(X_i'\beta_0)-g(X_i'\beta)-g^{(1)}(X_i'\beta)E[X_i'(\beta_0-\beta)\mid X_i'\beta] \\ &\quad +O(n^{-1}) \\ &= g^{(1)}(X_i'\beta)[X_i-E(X_i\mid X_i'\beta)](\beta_0-\beta)+O(n^{(-1)}) \end{aligned} \tag{8.44}$$

276

将式(8.44)代入 $S(\beta)=\sum_{i=1}^{n}[g(X_i'\beta_0)+u_i-G(X_i'\beta_0)]^2$ 我们得到

$$S(\beta)=(\beta_0-\beta)'W_0(\beta_0-\beta)-2V_0(\beta_0-\beta)+\sum_{i=1}^{n}u_i^2+o(1) \tag{8.45}$$

其中 W_0 与 V_0 均在式(8.10)中定义。由式(8.45)的一阶条件可以得到式(8.10)。

可以证明式(8.9)中首项可以通过将 $\hat{E}(Y_i\mid X_i'\beta)$ 替换为$E[g(X_i'\beta_0)\mid X_i'\beta]$得到。一致地，在 $\beta\in\mathcal{B}_n$上，我们有

$$\begin{aligned} S_{1n}(\beta) &= \frac{1}{n}\sum_i\{g(X_i'\beta_0)-E[g(X_i'\beta_0)\mid X_i'\beta]\}^2 \\ &\quad +2\frac{1}{n}\sum_i u_i\{g(X_i'\beta_0)-E[g(X_i'\beta_0)\mid X_i'\beta]\} \\ &\quad +\{\text{独立于}\beta\text{的项}\}+o_p(n^{-1}) \end{aligned} \tag{8.46}$$

由于 $\hat{\beta}\in\mathcal{B}_n$我们有 $|\hat{\beta}-\beta_0|=O(n^{-1/2})$，对 $g(X_i'\beta_0)$ 在 $\beta_0=\beta$ 处进行 Taylor 展开(下式进行了两次展开)，有

$$\begin{aligned} g(X_i'\beta_0)-E[g(X_i'\beta_0)\mid X_i'\beta] &= g(X_i'\beta_0)-g(X_i'\beta) \\ &\quad -g^{(1)}(X_i'\beta)E(X_i'\mid X_i'\beta)(\beta_0-\beta)+O_p(n^{-1}) \\ &= g^{(1)}(X_i'\beta)\{X_i'-E(X_i'\mid X_i'\beta)\}(\beta_0-\beta)+O_p(n^{-1}) \end{aligned} \tag{8.47}$$

将式(8.47)代入式(8.46)，我们得到

$$\begin{aligned} S_{1n}(\beta) &= (\beta_0-\beta)'\left[\frac{1}{n}\sum_i(g_i^{(1)})^2v_iv_i'\right](\beta_0-\beta) \\ &\quad +2\frac{1}{n}\sum_i u_ig_i^{(1)}v'(\beta_0-\beta) \\ &\quad +\{\text{独立于}\beta\text{的项}\}+o_p(n^{-1}) \end{aligned} \tag{8.48}$$

其中 $g_i^{(1)}=g^{(1)}(X_i'\beta)$，$v_i=X_i-E(X_i\mid X_i'\beta)$。

对 β 最小化 $S_{1n}(\beta,)$ 忽略掉与 β 无关项及 $o_p(n^{-1})$ 项，我们得到一阶条件

$$2(\beta_0-\beta)\frac{1}{n}\sum_i(g_i^{(1)})^2 v_i v_i' - 2\frac{1}{n}\sum_i u_i g_i^{(1)} v_i = 0 \tag{8.49}$$

277 由此可以导出

$$\sqrt{n}(\beta_0-\beta) = \left[\frac{1}{n}\sum_i(g_i^{(1)})^2 v_i v_i'\right]^{-1}\frac{1}{\sqrt{n}}\sum_i u_i g_i^{(1)} v_i$$
$$= \left[\frac{1}{n}\sum_i(g_{i0}^{(1)})^2 v_{i0} v_{i0}'\right]^{-1}\frac{1}{\sqrt{n}}\sum_i u_i g_{i0}^{(1)} v_{i0} + o_p(1) \tag{8.50}$$

其中 $g_{i0}^{(1)}=g^{(1)}(X_i'\beta_0)$，$v_{i0}=X_i-E(X_i|X_i'\beta_0)$，式(8.50)中第二个等式应用了 $\hat{\beta}-\beta_0=O_p(n^{-1/2})$ 的要求。

由标准的大数定律和 Lindeberg-Levy 中心极限定理可以推导出

$$\sqrt{n}(\beta_0-\beta)\to N(0,V_0)\text{ 依分布收敛} \tag{8.51}$$

其中

$$V_0=\Omega_0^{-1}\Sigma_0\Omega^{-1}$$
$$\Omega_0=E\{(g_{i0}^{(1)})^2 v_{i0}v_{i0}'\}$$
$$\Sigma_0=E\{\sigma^2(X_i)(g_{i0}^{(1)})^2 v_{i0}v_{i0}'\}$$
$$v_{i0}=X_i-E(X_i\mid X_i'\beta_0)$$

如果替换为 $w(X_i)=1$，定理 8.1 的结果与式(8.51)是相吻合的。

8.13 应　用

8.13.1 对邮寄目录直销响应的建模

直销通常用于这样的目标客户，基于可观测的特征如人口统计资料及个体历史购买决策，他们最可能会是回头客。例如，人们可能考虑只向那些最可能成为回头客或者“最像”回头客的人进行目录邮寄。① 然而，直销的成功或者失败，直接依赖于识别哪些客户最可能做出购买的能力。278

Racine(2002)考虑了一个行业标准数据库，这个数据库来自直销协会(DMA)②，它包含了一个复制礼品目录公司，即“每年向它的顾客邮寄一般及专门的目录若干次的高档礼品企业”的数据。基本时期包含 1971 年 12 月到 1992

① Bult 和 Wansbeek(1995)在最大化收益的框架下，指出事实上可能恰恰相反，卖方为了节省成本会避免向那些高度可能的购买者重复邮寄。不考虑目的的话，识别哪些人最可能进行购买的能力在过去已证明是个问题。

② 此数据库包含全美公开目录和非营利营销行业数据库大约 100 000 名顾客的购买历史。

年 6 月。DMA 收集的数据包括订单、14 个产品组中每组的购买量、购买时间以及购买方式。在 1992 年秋季较早时候向现有数据库中顾客至少发送了一份目录之后，数据就出现了三个月的缺口。然后从 1992 年 9 月到 1992 年 12 月的数据进行了更新。这给出了这样一个背景，模型可以在基本时期构建然后在后面的时期评估。在第一个时期构造出一个 4 500 名顾客的随机子集，为预测一个顾客购买可能性的各种方法被使用，接着对独立的后续(hold-out)样本来评估预测的准确性，样子包含从后面时期抽取的 1500 个随机挑选的个体。

对容量 $n_1 = 4\,500$ 的估计样本参数指数模型(Logit, Probit)和半参数指标模型(Ichimura(1993); Ichimura 和 Lee(1991))进行拟合，再对容量 $n_2 = 1\,500$ 的后续样本进行评估。

数据描述

我们有两个相互独立的估计数据集和评测数据集，容量分别为 $n_1 = 4\,500$ 和 $n_2 = 1\,500$，对每个顾客均有一份记录。我们把注意力放在其中一个产品组上，即选取 14 个产品组中部的第 8 个产品组。研究中涉及的变量如下，其中这些变量的特征见表 8.1 和表 8.2。

(i) Response：决定是否购买

(ii) LTDFallOrders：累计秋季订单

(iii) LastPurchSeason：购买发生季节[①]

(iv) Orders4YrsAgo：最近五年中的订单

(v) LTDPurchGrp8：累计购买

(vi) DateLastPurch：上次购买发生时间[②]

279

表 8.1　估计数据集一览($n_1 = 4\,500$)

变量	均值	标准误	最小值	最大值
Response	0.09	0.28	0	1
LTDFallOrders	1.36	1.38	0	15
LastPurchSeason	1.62	0.53	-1	2
Orders4YrsAgo	0.26	0.55	0	5
LTDPurchGrp8	0.09	0.31	0	4
DateLastPurch	37.31	27.34	0	117

① 在数据库中，当购买发生在 1 月至 6 月的此项记为 1，购买发生在 7 月至 12 月的记为 2，若没有购买发生记为 -1。

② 12/17 记为 0，1/72 记为 1，以此类推。

表 8.2　评测数据集一览($n_2 = 1\,500$)

变量	均值	标准误	最小值	最大值
Response	0.08	0.27	0	1
LTDFallOrders	1.32	1.38	0	14
LastPurchSeason	1.63	0.51	-1	2
Orders4YrsAgo	0.25	0.52	0	4
LTDPurchGrp8	0.08	0.29	0	3
DateLastPurch	36.44	26.95	0	116

每个模型都用其样本外表现来评测，样本外表现是基于 McFadden，Puig 和 Kirschner(1977)[①]的标准与正确购买分类率[②]。Logit 模型[③]的结果以混淆矩阵的形式在表 8.3 中给出，半参数指标模型的结果在表 8.4 中给出。

280

表 8.3　Logit 模型的混淆矩阵与分类率

	预测未购买	预测购买
实际未购买	1 378	5
实际购买	108	9
预测表现		91.95%
总正确分类率		92.47%
正确未购买分类率		99.64%
正确购买分类率		7.69%

表 8.4　半参数指标模型的混淆矩阵与分类率

	预测未购买	预测购买
实际未购买	1 361	22
实际购买	75	42
预测表现		93.26%
总正确分类率		93.53%
正确未购买分类率		98.41%
正确购买分类率		35.90%

① $P_{11} + P_{22} - P_{21}^2 - P_{12}^2$，其中 P_{ij} 为 2×2 混淆矩阵中第 i 行第 j 列元素，表示为所有元素加总后的一部分。“混淆矩阵”是指模型真实结果与预测结果对比的一个列表。对角线元素包含正确的预测结果而非对角线元素包含不正确(混淆)的结果。

② 最终确实购买的顾客中被准确预测到的所占的比例。

③ Probit 模型的结果没有 Logit 模型的结果好，为节省空间将其略去。

半参数单指标模型对后续数据得出的预测表现要优于参数 Logit 模型。另外注意，尽管在 McFadden 等(1977)的标准下，参数模型看起来表现不错，但是在对确实做出购买的这类人的预测上表现不佳。

8.14 习　题

281

习题 8.1

(i) 如果 Y 为一个 $\{0,1\}$ 二元变量，证明 $P(Y=1|x)=E(Y|x)$。

(ii) 如果 Y 为一个在 $\{1,2\}$ 中取值的二元变量，$P(Y=1|x)$ 与 $E(Y|x)$ 仍然相等吗？

习题 8.2　在推导式(8.51)时，对式(8.51)右端第二项，我们应用了如下事实

(i) $\frac{1}{n}\sum_i u_i\hat{E}(u_i \mid X_i'\beta)$ = 独立于 $\beta + o_p(n^{-1})$ 的项

(ii) $\frac{1}{n}\sum_i u_i\{E[g(X_i'\beta_0)] - \hat{E}[g(X_i'\beta_0) \mid X_i'\beta]\} = o_p(n^{-1})$，其中

$$\hat{E}(u_i \mid X_i'\beta) = (nh)^{-1}\sum_{j\neq i} u_jK(((X_i - X_j)'\beta/h)/\hat{p}(X_i'\beta)$$

$$\hat{p}(X_i'\beta) = (nh)^{-1}\sum_{j\neq i} K(((X_i - X_j)'\beta/h)$$

证明(i)和(ii)。

提示：对(i)，在 $\beta=\beta_0$ 处进行 Taylor 展开并应用 $\beta-\beta_0=O_p(n^{-1/2})$。Taylor 展开式中第一项是独立于 β 的。第二项则是一个二阶 U 统计量，应用 H 分解可知其为 $o_p(n^{-1})$ 阶的。

习题 8.3　证明

$$\frac{1}{n}\sum_i [g(X_i'\beta_0) - \hat{E}(Y_i \mid X_i'\beta)]^2 = \frac{1}{n}\sum_i [g(X_i'\beta_0) - E(g(X_i'\beta_0) \mid X_i'\beta)]^2 + \text{独立于 } \beta + o_p(n^{-1}) \text{ 的项}$$

提示：记

$$\begin{aligned}\hat{E}(Y_i \mid X_i'\beta) &= \hat{E}(g(X_i'\beta_0) \mid X_i'\beta) + \hat{E}(u_i \mid X_i'\beta) \\ &= E(g(X_i'\beta_0) \mid X_i'\beta) + [\hat{E}(g(X_i'\beta_0) \mid X_i'\beta) - E(g(X_i'\beta_0) \mid X_i'\beta)] \\ &\quad + \hat{E}(u_i \mid X_i'\beta)\end{aligned}$$

其中 $\hat{E}(g(X_i'\beta_0)|X_i'\beta)$ 与 $\hat{E}(u_i|X_i'\beta)$ 的定义见 8.12 节。

习题 8.4　验证当应用 Ai(1997)一般方法于二元变量 y 的单指标模型时，Ai 的一阶条件与式(8.52)中 Klein 和 Spady(1993)的一阶条件相吻合。

282 提示：Klein 和 Spady(1993)的估计量解出了一阶条件(见 Pagan 和 Ullah (1999, p. 283)) $\sum_{i=1}^{n}\hat{m}_i(\theta)=0$，其中

$$\hat{m}_i(\theta)=\hat{g}(x'\theta)^{-1}[1-\hat{g}(X_i'\theta)]^{-1}\frac{\partial\hat{g}(X_i'\theta)}{\partial\theta}[Y_i-\hat{g}(X_i'\theta)] \quad (8.52)$$

其中 $\hat{g}(X_i'\theta)=\hat{E}(Y_i\mid X_i'\theta)=\dfrac{\sum_{j=1}^{n}Y_jK\left(\dfrac{(X_j-X_i)'\theta}{h}\right)}{\sum_{j=1}^{n}K\left(\dfrac{(X_j-X_i)'\theta}{h}\right)}$。

第9章 可加和平滑(变)系数半参数模型

在本章中我们考虑文献中出现的一些流行的半参数回归模型。应用这些半参数模型而非纯粹的非参数模型的基本原因在于,半参数模型可以减少非参数部分的维数,从而减轻维数诅咒。当然,这些模型也可能像纯粹的参数模型一样受到函数形式误设的批评,但是这些模型已被证明在应用设定方面极其流行,且在解释方面比纯粹非参数模型更趋于简单。 283

9.1 一个可加模型

我们首先考虑半参数可加模型

$$Y_i = c_0 + g_1(Z_{1i}) + g_2(Z_{2i}) + \cdots + g_q(Z_{qi}) + u_i \tag{9.1}$$

其中 c_0 为一个标量参数,Z_{li} 均为一元连续变量,$g_l(\cdot)$ $(l=1,\cdots,q)$ 是未知的平滑函数。观测值 $\{Y_i, Z_{1i}, \cdots, Z_{qi}\}_{i=1}^{n}$ 被假定是独立同分布的。

对于基于核的方法,两种方法在估计可加模型时最常用:"反向拟合"(backfitting)法(见 Bujia, Hastie 和 Tibshirani(1989);Hastie 和 Tibshirani(1990))与"边际积分"法,后者由 Linton 和 Nielsen(1995)、Newey(1994b) 284 以及 Tjøstheim 和 Auestad(1994)分别独立提出。另外也可参考 Chen, Härdle, Linton 和 Severance-Lossin(1996)以及 Linton(1997,2000)。由于其迭代性质,反向拟合法比边际积分法分析起来要困难许多,因此我们先讨论相对比较简单的边际积分法。

9.1.1 边际积分法

首先,注意对任意常数 c,有$[g_1(z_1)+c]+[g_2(z_2)$

$-c]=g_1(z_1)+g_2(z_2)$。因此为了识别函数 $g_l(\cdot)$,我们需要一些识别条件。由核方法很方便施加条件$E[g_l(Z_l)]=0$,于是 $g_l(\cdot)$的各个元素可被识别($l=1,\cdots,q$)。这也导致$E(Y_i)=c_0$。

令 $Z_{\underline{\alpha}i}=(Z_{1i},\cdots,Z_{\alpha-1,i},Z_{\alpha+1,i},\cdots,Z_{qi})$,即 $Z_{\underline{\alpha}i}$是通过$(Z_{1i},Z_{2i},\cdots,Z_{qi})$移去 $Z_{\alpha i}$ 得到的。定义 $G_{\underline{\alpha}}(z_{\underline{\alpha}})=g_1(z_1)+\cdots+g_{\alpha-1}(z_{\alpha-1})+g_{\alpha+1}(z_{\alpha+1})+\cdots g_q(z_q)$。用这些记号,则式(9.1)可以写成

$$Y_i = c_0 + g_\alpha(Z_{\alpha i}) + G_{\underline{\alpha}}(Z_{\underline{\alpha}i}) + u_i. \tag{9.2}$$

定义 $\xi(z_\alpha,Z_{\underline{\alpha}j})=E(Y_j|Z_{\alpha j}=z_\alpha,Z_{\underline{\alpha}j})$。对式(9.2)两端应用 $E[\cdot|Z_{\alpha j}=z_\alpha,Z_{\underline{\alpha}j}]$,我们得到

$$\xi(z_\alpha,Z_{\underline{\alpha}j}) = c_0 + g_\alpha(z_\alpha) + G_{\underline{\alpha}}(Z_{\underline{\alpha}j}). \tag{9.3}$$

定义 $m_\alpha(z_\alpha)=E[\xi(z_\alpha,Z_{\underline{\alpha}j})]$,注意这里是关于 $Z_{\underline{\alpha}j}$ 取边际期望的。对式(9.3)取期望(关于 $Z_{\underline{\alpha}j}$),我们得到

$$m_\alpha(z_\alpha) = c_0 + g_\alpha(z_\alpha) \tag{9.4}$$

其中我们用到了 $E[G_{\underline{\alpha}}(Z_{\underline{\alpha}j})]=0$。由式(9.4)可得

$$g_\alpha(z_\alpha) = m_\alpha(z_\alpha) - E[m_\alpha(Z_{\alpha i})] \tag{9.5}$$

上述方法是不可行的。然而,一个可行的估计量可以通过将期望替换为样本均值,边际期望(积分)替换为边际平均值,条件均值函数替换为局部线性核估计量来得到。特别地,式(9.5)一个可行的对应估计量为

$$\tilde{g}_\alpha(z_\alpha) = \tilde{m}_\alpha(z_\alpha) - n^{-1}\sum_{i=1}^{n}\tilde{m}_\alpha(Z_{\alpha i}) \tag{9.6}$$

285 其中

$$\tilde{m}_\alpha(z_\alpha) = n^{-1}\sum_{j=i}^{n}\tilde{\alpha}_\alpha(Z_\alpha,Z_{\underline{\alpha}j}) \tag{9.7}$$

且 $\tilde{a}_\alpha(z_\alpha,Z_{\underline{\alpha}j})$为下述最小化问题中 a 的解,

$$\min_{\{a,b\}}\sum_{l=1}^{n}[y_l - a - (Z_{\alpha l}-z_\alpha)b]^2 k_{h_\alpha}(Z_{\alpha l}-z_\alpha)K_{h_{\underline{\alpha}}}(Z_{\underline{\alpha}l}-Z_{\underline{\alpha}j})。$$

注意 $\tilde{a}_\alpha=(z_\alpha,Z_{\underline{\alpha}j})$是 $E(Y_j|Z_{\alpha j}=z_\alpha,Z_{\underline{\alpha}j})$的核估计量,它仅对 z_α 运用局部线性方法,而对 $z_{\underline{\alpha}}$ 只运用局部常数方法。

Fan, Härdle 和 Mammen(1998)假定 $k(\cdot)$为一个一元二阶核,$K(\cdot)$为 v 阶乘积核,其中 v 是满足 $v\geq(q-1)/2$ 的正整数。因此,若 $q>4$,一个高阶核就是必需的($v>2$);有关高阶核函数的构造见 1.11 节。下述定理给出了 $\tilde{g}_\alpha(z_\alpha)$的渐近分布,为了表述简单,我们假定有 $h_1=\cdots=h_{\alpha-1}=h_{\alpha+1}=\cdots=h_q=h_{\underline{\alpha}}$。Fan 等证明了如下结果。

定理 9.1 在 Fan 等(1998)给出的条件下,另假定 $nh_\alpha h_{\underline{\alpha}}^{q-1}/\ln n\to\infty$,

$h_{\underline{\alpha}}^{v}/h_{\alpha}^{2}\to 0$, $h_{\alpha}\to 0$ 和 $h_{\underline{\alpha}}\to 0$,那么

$$\sqrt{nh_{\alpha}}\left\{\tilde{g}_{\alpha}(z_{\alpha}) - g_{\alpha}(z_{\alpha}) - c_{0} - \frac{1}{2}h_{\alpha}^{2}\kappa_{2}g_{\alpha}^{(2)}(z_{\alpha}) + o(h_{\alpha}^{2})\right\} \\ \to N(0,v_{\alpha}(z_{\alpha}))$$

其中 $g_{\alpha}^{(2)}(\cdot)$ 是 $g_{\alpha}(\cdot)$ 的二阶导数, $\kappa_{2} = \int u^{2}k(u)\mathrm{d}u$,

$$v_{\alpha}(z_{\alpha}) = f_{\alpha}(z_{\alpha})\left[\int k^{2}(u)\mathrm{d}u\right]E\left\{\frac{\sigma^{2}(z_{\alpha},Z_{\underline{\alpha}i})f_{\underline{\alpha}}^{2}(z_{\underline{\alpha}})}{f^{2}(z_{\alpha},z_{\underline{\alpha}})} \mid Z_{\alpha i} = z_{\alpha}\right\}$$

且 $\sigma^{2}(z_{\alpha},z_{\underline{\alpha}}) = E(u_{i}^{2} \mid Z_{\alpha i} = z_{\alpha},Z_{\underline{\alpha}i} = z_{\underline{\alpha}})$。

定理9.1证明了 $\tilde{g}(z_{\alpha})$ 达到了一维最优收敛速度。

边际积分方法的一个缺点是计算量较大。人们不得不对所有 $i,j = 1,\cdots n$ 来估计 $E[Y_{l} \mid Z_{\alpha l} = Z_{\alpha i},Z_{\underline{\alpha}l} = Z_{\underline{\alpha}j}]$,这是一个 n^{2} 级的计算,而估计非可加模型相应只需要 n 级。在下节我们讨论一个在计算上有效率的估计可加模型的方法。 286

9.1.2　一个计算上有效的 oracle 估计量

上节讨论的边际积分方法由于需要对 $i,j = 1,\cdots,n$ 估计 $g(Z_{\alpha i},Z_{\underline{\alpha}j})$,故计算量非常大。Kim,Linton 和 Hengartner(1999)提出了一个备选方法,可以将估计时间消减至 n 级。考虑对 Y 在 Z_{α} 上的非参数回归,

$$E(Y \mid Z_{\alpha} = z_{\alpha}) = c_{0} + g_{\alpha}(z_{\alpha}) + \sum_{s\neq\alpha}^{q}E[g_{s}(Z_{s}) \mid Z_{\alpha} = z_{\alpha}] \tag{9.8}$$

由于 $\sum_{s\neq\alpha}E[g_{s}(Z_{s}) \mid Z_{\alpha} = z_{\alpha}]$ 的存在,可证明 $E(Y \mid Z_{\alpha} = z_{\alpha})$ 是 $c_{0} + g_{\alpha}(z_{\alpha})$ 的一个有偏估计量。Kim 等建议选择一个工具变量 $w_{\alpha}(z)$ 使得

$$E[w_{\alpha}(Z) \mid Z_{\alpha} = z_{\alpha}] = 1$$

$$E[w_{\alpha}(Z)g_{s}(X_{s}) \mid Z_{\alpha} = z_{\alpha}] = 0, s \neq \alpha \tag{9.9}$$

那么可以得到 $E[w_{\alpha}(Z_{i})Y_{i} \mid Z_{\alpha i} = z_{\alpha}] = c_{0} + g_{\alpha}(z_{\alpha})$。容易验证 $w_{\alpha}(z) = f_{\alpha}(z_{\alpha})f_{\underline{\alpha}}(z_{\underline{\alpha}})/f(z_{\alpha},z_{\underline{\alpha}})$ 就是一个满足式(9.9)的函数,其中 $f_{\underline{\alpha}}(z_{\underline{\alpha}})$ 是 $z_{\underline{\alpha}} = (z_{1},\cdots,z_{\alpha-1},z_{\alpha+1},\cdots,z_{q})$ 的联合概率密度函数。事实上,对任意随机变量 ξ,我们有

$$\begin{aligned} E[\xi w_{\alpha}(z) \mid z_{\alpha}] &= \int \xi w_{\alpha}(z)\frac{f(z_{\alpha},z_{\underline{\alpha}})}{f_{\alpha}(z_{\alpha})}\mathrm{d}z_{\underline{\alpha}} \\ &= \int \xi\frac{f_{\alpha}(z_{\alpha})f_{\underline{\alpha}}(z_{\underline{\alpha}})}{f(z_{\alpha},z_{\underline{\alpha}})}\frac{f(z_{\alpha},z_{\underline{\alpha}})}{f_{\alpha}(z_{\alpha})}\mathrm{d}z_{\underline{\alpha}} \\ &= \int \xi f_{\underline{\alpha}}(z_{\underline{\alpha}})\mathrm{d}z_{\underline{\alpha}} \end{aligned} \tag{9.10}$$

它正好是 ξ 关于 $z_{\underline{\alpha}}$ 部分的边际积分。将式(9.10)中的 ξ 替换为 $\xi = 1$ 可以得到

$E[w_\alpha(Z) \mid Z_{\alpha i}] = \int f_{\underline{\alpha}}(z_{\underline{\alpha}})\mathrm{d}z_{\underline{\alpha}} = 1$。同样,对于 $s \neq \alpha$,将 ξ 替换为 $g_s(Z_{si})$,那么有 $E[w_\alpha(Z_i)g_s(Z_{si}) \mid Z_{\alpha i}] = \int g_s(Z_{si})f_{\underline{\alpha}}(z_{\underline{\alpha}})\mathrm{d}z_{\underline{\alpha}} = E[g_s(Z_{si})] = 0$。于是式(9.9)成立。

287 容易知道

$$\tilde{\gamma}_\alpha(z_\alpha, h_o) = \frac{1}{nh_o}\sum_{j=1}^{n} k\left(\frac{Z_{\alpha j} - z_\alpha}{h_o}\right)\frac{\hat{f}_{\underline{\alpha}}(Z_{\underline{\alpha} j})}{\hat{f}(Z_{\alpha j}, Z_{\underline{\alpha} j})}Y_j$$

$$= \frac{1}{n^2 h_o^q}\sum_{j=1}^{n}\sum_{l=1}^{n}\frac{k\left(\frac{z_{\alpha j} - z_\alpha}{h_o}\right)k_{\underline{\alpha}}\left(\frac{z_{\underline{\alpha} l} - z_{\underline{\alpha} j}}{h_o}\right)}{\hat{f}(z_{\alpha j}, Z_{\underline{\alpha} j})}Y_j \qquad (9.11)$$

为 $E[w_\alpha(Z)Y \mid Z_\alpha = z_\alpha] = c_0 + g_\alpha(z_\alpha)$ 的一个一致估计量,其中 $w_\alpha(z) = f_\alpha(z_\alpha)f_{\underline{\alpha}}(z_{\underline{\alpha}})/f(z_\alpha, z_{\underline{\alpha}})$。

注意 $\tilde{\gamma}_\alpha(z_\alpha)$ 可以解释为对修正过的 $\hat{Y}_{\alpha j}$ 关于 $Z_{\alpha j}$ 进行回归得到的一个一维标准局部常数估计量,其中 $\hat{Y}_{\alpha j} = Y_j\hat{f}_\alpha(z_\alpha)\hat{f}_{\underline{\alpha}}(Z_{\underline{\alpha} j})/\hat{f}(Z_{\alpha j}, Z_{\underline{\alpha} j})$。由 $\tilde{\gamma}_\alpha(z_\alpha)$ 估计 $c_0 + g_\alpha(z_\alpha)$,自然地可以用 $\tilde{g}_\alpha(z_\alpha) = \tilde{\gamma}_\alpha(z_\alpha) + n^{-1}\sum_{j=1}^{n}\tilde{\gamma}_\alpha(z_{\alpha j})$ 来估计 $g_\alpha(z_\alpha)$。Kim 等(1999)中定理 1 给出了 $\tilde{g}_\alpha(z_\alpha)$ 的偏误首项、方差误和渐近正态分布。Kim 等证明了 $\tilde{g}_\alpha(z_\alpha)$ 的渐近方差比 $g_\alpha(z_\alpha)$ 的边际积分估计量的渐近方差多了一项。因此,$\tilde{g}_\alpha(z_\alpha)$ 与边际积分估计量相比是弱有效的。Kim 等进一步提出了一个有效的 oracle 估计量,下面我们将对其进行讨论。

$g_\alpha(z_\alpha)$ 的一个 oracle(有效的)估计量被定义为,它具有与其他所有可加函数均已知情形下相同的偏误首项和方差首项。定义 $Y_{\alpha i}^{\text{oracle}} = Y_i - \sum_{s \neq \alpha} g_s(Z_{si}) - c_0$,考虑 $g_\alpha(z_\alpha)$ 的一个使用了 $\{Z_{\alpha i}, Y_{\alpha i}^{\text{oracle}}\}_{i=1}^{n}$ 的一元核估计量,即,

$$\tilde{g}_\alpha^{\text{oracle}}(z_\alpha) = \frac{\sum_{j=1}^{n} k\left(\frac{z_{\alpha j} - z_\alpha}{h_\alpha}\right)Y_{\alpha j}^{\text{oracle}}}{\sum_{j=1}^{n} k\left(\frac{z_{\alpha j} - z_\alpha}{h_\alpha}\right)} \qquad (9.12)$$

其中 $h_\alpha = \delta_1 n^{-1/5}$($\delta > 0$ 为一常数)。应用第 2 章得到的结果,我们知道(使用一个二阶核函数)

$$\sqrt{nh_\alpha}\{\tilde{g}_\alpha^{\text{oracle}}(z_\alpha) - g_\alpha(z_\alpha) - h_\alpha^2 b_\alpha(z_\alpha)\} \to N(0, V_\alpha(z_\alpha)) \text{ 依分布收敛} \qquad (9.13)$$

其中 $b(z_\alpha) = \kappa_2\left[\frac{1}{2}g_\alpha^{(2)}(z_\alpha) + g_\alpha^{(1)}(z_\alpha)f_\alpha^{(1)}(z_\alpha)\right]/f_\alpha(z_\alpha)$,而 $V_\alpha(z_\alpha) = \kappa\sigma_\alpha^2(z_\alpha)/$

$f_\alpha(z_\alpha)(\sigma_\alpha^2(z_\alpha) = \mathrm{var}(Y|Z_\alpha = z_\alpha))$。

可以使用一个局部线性估计量而非局部常数估计量来估计 $g_\alpha(z_\alpha)$。也就是说,令 $\tilde{a}$ 和 $\tilde{b}$ 为最小化以下目标函数的 a 和 b 的值: 288

$$\sum_{i=1}^{n} k\left(\frac{Z_{\alpha i} - z_\alpha}{h_\alpha}\right)[Y_{\alpha i}^{\text{oracle}} - a - b(Z_{\alpha i} - z_\alpha)]^2$$

那么,$\tilde{a}$ 就是 $g_\alpha(z_\alpha)$ 的局部线性 oracle 估计,它的渐近分布与第 2 章所给出的基本相同,除了将式(9.13)中偏误首项变为 $(1/2)\kappa_2 g_\alpha^{(2)}(z_\alpha)/f_\alpha(z_\alpha)$。

以上给出的估计量 $\tilde{g}_\alpha^{\text{oracle}}(z_\alpha)$ 是非可行的,这是由于 $s \neq \alpha$ 时,c_0 与 $g_s(z_s)$ 都是未知的。我们可以用 $\tilde{\gamma}_s(z_s)$ 替换 $g_s(z_s)$,用 $\bar{Y}$ 替换 c_0。于是我们可以用

$$Y_{\alpha i}^{\text{2-step}} = Y_i - \sum_{s \neq \alpha}^{q} \tilde{\gamma}_s(Z_{si}, h_\alpha) + (q-1)\bar{Y}$$

来替换 $Y_{\alpha i}^{\text{oracle}} = Y_i - \sum_{s \neq \alpha}^{q} g_s(Z_{si}) - c_0$,其中

$$\tilde{\gamma}_s(z_s, h_\alpha) = \frac{1}{nh_\alpha}\sum_{j=1}^{n} k\left(\frac{Z_{sj} - z_s}{h_\alpha}\right)\frac{\hat{f}_{\underline{s}}(Z_{\underline{s}j})}{\hat{f}(Z_{sj}, Z_{\underline{s}j})}Y_j \tag{9.14}$$

是 $E[w_s(Z)Y|Z_s = z_s] = c_0 + g_s(z_s)$ 的一个一致估计量,$s = 1, \cdots, q$。

定义 $\hat{g}_\alpha^{\text{oracle}}$ 为式(9.12)中用 $Y_{\alpha i}^{\text{2-step}}$ 替换 $Y_{\alpha i}^{\text{oracle}}$ 后的 $\tilde{g}_\alpha^{\text{oracle}}$,Kim 等(1999)证明了 $\hat{g}_\alpha^{\text{oralee}}$ 具有与 $\tilde{g}_\alpha^{\text{oracle}}$ 同样的偏误首项和方差首项;因此,它也有同样的渐近分布。Kim 等假定对某些 $\alpha \geqslant 0$,$h_\alpha = an^{-1/5}$,$h_o = (n^{-1/5})$。注意这里选择的 h_o 是比 h_α 更低阶的。他们证明,若使用一个局部线性估计量,那么

$$\sqrt{nh_\alpha}\left\{\hat{g}_\alpha^{\text{oracle}}(z_\alpha) - g_\alpha(z_\alpha) - \frac{\kappa_2}{2}h_\alpha^2 g_\alpha^{(2)}(z_\alpha)/f_\alpha(z_\alpha)\right\} \xrightarrow{d} N(0, V_\alpha(z_\alpha)) \text{ 依分布收敛} \tag{9.15}$$

其中 $V_\alpha(z_\alpha)$ 的定义见式(9.13)下方。

式(9.15)表明可行估计量 $\hat{g}_\alpha^{\text{oralee}}$ 具有与所有其他 $g_s(z_s)$ 均已知情形相同的一阶有效性($s \neq \alpha$)。

Kim 等(1999)进一步证明了一个原始自举法(wild bootstrap procedure)可以用来更好地逼近 $\hat{g}_\alpha^{\text{oralee}}(\cdot)$ 的有限样本分布。定义一个估计出的可加函数 289

$$\hat{g}_{\text{add}}(z; h_\alpha, h_0) = \bar{Y} + \sum_{\alpha=1}^{q} \hat{g}_\alpha(z_\alpha; h_\alpha, h_o),$$

其中 $\hat{g}_\alpha(z_\alpha; h_\alpha, h_o) = \hat{g}_\alpha(z_\alpha)$ 是 $g_\alpha(z_\alpha)$ 初始的可加函数估计量。这样 u_i 可以用 $\hat{u}_i = Y_i - \hat{g}_{\text{add}}(Z_i; h_\alpha, h_o)$ 来估计。定义中心化残差(centered residual)为 $\tilde{u} = \hat{u}_i -$

$n^{-1}\sum_{l=1}^{n}\hat{u}_l$。则自举法误差 u_i^* 可由两点原始自举法[①]得到，即 $u_i^*=[(1+\sqrt{5})/2]\tilde{u}_i$ 的概率为 $r=(1+\sqrt{5})/(2\sqrt{5})$，$u_i^*=[(1-\sqrt{5})/2]u_i$ 的概率为 $1-r$。接下来有

$$Y_i^* = \hat{g}_{\text{add}}(Z_i;\bar{h}_\alpha,h_o) + u_i^*$$

其中使用了另外一个窗宽 $\bar{h}_\alpha$。Kim 等证明为了使自举法可用，$\bar{h}_\alpha$ 的阶数要大于 h_α。Kim 等建议选择 $h_\alpha \sim n^{-1/5}$，$h_o=o(h_\alpha)$，并且令 $\bar{h}_\alpha$ 在 $[n^{-1/5+\delta},n^{-\delta}]$ 中取值，其中 $0<\delta<1/5$。这样就可以用自举样本 $\{Z_i,Y_i\}_{i=1}^n$ 来计算 $\hat{g}_\alpha^{*,\text{oracle}}(z_\alpha)$，其中 $\hat{g}_\alpha^{*,\text{oracle}}(z_\alpha)$ 与 $\hat{g}_\alpha^{\text{oracle}}(z_\alpha)$ 相同，除了将所有 Y_i 替换为 Y_i^*。Kim 等举例证明可以重复生成 $\hat{g}_\alpha^{*,\text{oracle}}(Z_\alpha)$ 的 B 自举估计并用他们的经验分布来逼近 $\hat{g}_\alpha^{\text{oracle}}(z_\alpha)$ 的有限样本分布。

9.1.3 普通反向拟合法

在 $Z_{\alpha i}=z_\alpha$ 的条件下对式(9.1)取条件期望，我们得到($a=1,\cdots,q$)

$$g_\alpha(z_\alpha) = E(Y_i \mid Z_{\alpha i}=z_\alpha) - c_0 - \sum_{s\neq\alpha}^{q} E[g_s(Z_{si}) \mid Z_{\alpha i}=z_\alpha] \quad (9.16)$$

式(9.16)显示迭代过程是恰当的。令 $\hat{g}_\alpha^{[0]}(z_\alpha)$ 为 $g_\alpha(z_\alpha)$ 的某个初始值，如

290

$\hat{g}_\alpha^{[0]}(z_\alpha)=0$ 或者 $\hat{g}_\alpha^{[0]}(z_\alpha)$ 是 $g_\alpha(z_\alpha)$ 的边际积分估计量。同样，$\hat{c}_0=n^{-1}\sum_{i=1}^{n}Y_i$。那么迭代过程由下式给出。对 $\alpha=1,\cdots,q$，$l=1,2,3,\cdots$，用下式来计算第 l 步 $\hat{g}_\alpha^{[l]}(z_\alpha)$：

$$\hat{g}_\alpha^{[l]}(z_\alpha) = \hat{E}(Y_i \mid Z_{\alpha i}=z_\alpha) - \hat{c}_0 - \sum_{s=1}^{\alpha-1}\hat{E}[\hat{g}_s^{[l]}(Z_{si}) \mid Z_{\alpha i}=z_\alpha]$$
$$- \sum_{s=\alpha+1}^{q}\hat{E}[\hat{g}_s^{[l-1]}(Z_{si}) \mid Z_{\alpha i}=z_\alpha] \quad (9.17)$$

其中，对随机变量 A_i，$\hat{E}[A_i \mid Z_{\alpha i}=z\alpha]$ 是 $E[A_i \mid Z_{\alpha i}=z_\alpha$ 的(一元)核估计量。$\hat{E}[A_i|Z_{\alpha i}=z_\alpha]$ 可以是局部常数估计量 $\sum_{j=i}^{n}A_jk_{\alpha,z,z_j}/\sum_{j=1}^{n}k_{\alpha,z,z_j}$，其中 $k_{\alpha,z,z_j}=k((Z_{\alpha j}-z_\alpha)/h_\alpha)$，或者也可以是 $E[A_i \mid Z_{\alpha i}=z_\alpha]$ 的局部线性估计量。

当预定的收敛标准达到时，例如，

$$\sum_{i=1}^{n}[\hat{g}_\alpha^{[l]}(Z_{\alpha i}) - \hat{g}_\alpha^{[l-1]}(Z_{\alpha i})]^2 \Big/ \sum_{i=1}^{n}(\hat{g}_\alpha^{[l-1]}(Z_{\alpha i}))^2$$

小于某一很小的数，如 10^{-4} 时，迭代过程终止。

---

① 更多关于原始自举法的详细讨论见第 12 章。

通过在上述过程中应用一个局部线性平滑子,Opsomer 和 Ruppert(1998)以及 Opsomer (2000)证明了估计的偏误是 $O(\sum_{\alpha=1}^{q} h_\alpha^2)$,且 $\hat{g}_\alpha(z_\alpha)$ 的方差首项如下($\kappa = \int k^2(v)\,\mathrm{d}v$)

$$\mathrm{var}(\hat{g}_\alpha(z_\alpha)) = \frac{\kappa\sigma^2}{nh_\alpha f_\alpha(x_\alpha)} + o\left(\frac{1}{nh_\alpha}\right)$$

其中 $\sigma^2 = E(u_i^2 | Z_i) = E(u_i^2)$(假定条件同方差)。

9.1.4 平滑反向拟合法

Mammen,Linton 和 Nielsen(1999)提出了一个平滑反向拟合过程来估计可加模型(9.1)。其思路是将 Y(或者 $E(Y|z)$)投射到可加函数空间上。令 $\tilde{g}(z) = \sum_{i=1}^{n} Y_i K_h(Z_i,z) / \sum_{i=1}^{n} K_h(Z_i,z)$ 和 $\tilde{f}(z) = n^{-1}\sum_{i=1}^{m} K_h(Z_i,z)$ 分别表示 $E(Y|Z=z)$ 和 $f(z)$ 的多维核估计量,其中 $z = z_i,\cdots,z_q$。$g_1(z_1),\cdots,g_q(z_q)$ 的局部常数平滑反向拟合估计量定义为那些使以下目标函数最小化的 $g_1,\cdots,g_q$:

$$\int[\tilde{g}(z) - g_1(z_1) - \cdots - g_q(z_q)]^2 \tilde{f}(z)\,\mathrm{d}z \tag{9.18}$$

291

其中最小化是关于所有满足 $\int g_\alpha(z_\alpha)\tilde{f}_\alpha(z_\alpha)\,\mathrm{d}z_\alpha = 0$ 的函数 $g(z_\alpha) = c_0 + \sum_{\alpha=1}^{q} g_\alpha(z_\alpha)$ 进行的,$\tilde{f}_\alpha(z_\alpha) = \int \tilde{f}(z)\,\mathrm{d}z_{\underline{\alpha}}$ 是边际概率密度函数$f_\alpha(z_\alpha)$的核估计量,$z_{\underline{\alpha}} = z_1\cdots,z_{\alpha-1},z_{\alpha+1},\cdots,z_q$。

式(9.18)的解可以由下列方程组来刻画($\alpha = 1,\cdots,q; \bar{Y} = n^{-1}\sum_{i=1}^{n} Y_i$);

$$\hat{g}_\alpha(z_\alpha) = \int\tilde{g}(z)\frac{\tilde{f}(z)}{\tilde{f}_\alpha(z_\alpha)}\mathrm{d}z_{\underline{\alpha}} - \sum_{s\neq\alpha}^{q}\int\hat{g}(z_s)\frac{\tilde{f}(z)}{\tilde{f}_\alpha(z_\alpha)}\mathrm{d}z_{\underline{\alpha}} - \bar{Y} \tag{9.19}$$

$$0 = \int\hat{g}_\alpha(z_\alpha)\tilde{f}_\alpha(z_\alpha)\,\mathrm{d}z_\alpha \tag{9.20}$$

注意

$$\int\tilde{g}(z)\frac{\tilde{f}(z)}{\tilde{f}_\alpha(z_\alpha)}\mathrm{d}z_{\underline{\alpha}} = \frac{(nh_\alpha)^{-1}\sum_{i=1}^{n} Y_i k(z_\alpha, Z_{\alpha i})}{\tilde{f}_\alpha(z_\alpha)} \equiv \tilde{g}_\alpha(z_\alpha) \tag{9.21}$$

这是由于 $\int\prod_{s\neq\alpha} h_s^{-1}k((z_s - Z_{si})/h_s)\,\mathrm{d}z_{\underline{\alpha}} = 1$,其中 $\tilde{g}_\alpha(z_\alpha)$ 是 $E(Y_i | Z_{\alpha i} = z_\alpha)$ 的局

部常数估计量。另外,应用 $\hat{g}_\alpha(z_\alpha)$ 的可加性及 $\int h_s^{-1}k((z_s - Z_{si})/h_s)\mathrm{d}z_s = 1$ 的事实,$q-1$ 重积分 $\int \mathrm{d}z_{\underline{\alpha}}$ 可以简化为二重积分,于是

$$\sum_{s\neq\alpha}\int \hat{g}_\alpha(z)\frac{\tilde{f}(z)}{\tilde{f}_\alpha(z_\alpha)}\mathrm{d}z_{\underline{\alpha}} = \sum_{s\neq\alpha}\int \hat{g}_\alpha(z)\frac{\tilde{f}_{\alpha,s}(z_\alpha,z_s)}{\tilde{f}_\alpha(z_\alpha)}\mathrm{d}z_s \tag{9.22}$$

其中 $\tilde{f}_{\alpha,s}(z_\alpha,z_s) = (nh_sh_\alpha)^{-1}\sum_{i=1}^{n}k((Z_{\alpha i}-z_\alpha)/h_\alpha)k((Z_{si}-z_s)/h_s)$ 是(z_α,z_s)的二维边际概率密度函数的核估计量。

由式(9.19)、式(9.21)和式(9.22)可得到如下的迭代过程:

$$\hat{g}_\alpha^{[r+1]}(z_\alpha) = \tilde{g}_\alpha(z_\alpha) - \sum_{s=1}^{\alpha-1}\int \hat{g}_s^{[r+1]}(z_s)\frac{\tilde{f}_{\alpha,s}(z_\alpha,z_s)\mathrm{d}z_s}{\tilde{f}_\alpha(z_\alpha)} - \sum_{s=\alpha+1}^{q}\int \hat{g}_s^{[r]}(z_s)\frac{\tilde{f}_{\alpha,s}(z_\alpha,z_s)\mathrm{d}z_s}{\tilde{f}_\alpha(z_\alpha)}$$

其中 $r=1,2,\cdots$,表示迭代的次数。

Mammen 等(1999)推导出了 $\hat{g}_\alpha(z_\alpha)$的渐近分布。为了表述他们的结果,对于一给定的函数$\beta(z)$,我们首先定义常数 β_0 和一元函数 $\beta_\alpha(z_\alpha)$($\int\beta_\alpha(z_\alpha)\mathrm{d}z_\alpha=0$, $\alpha=1,\cdots,q$)为

292

$$(\beta_0,\cdots,\beta_q) = \arg\min_{\beta_0,\cdots,\beta_q}\left[\int\{\beta(z)-\beta_0-\beta_1(z_1)-\cdots-\beta_q(z_q)\}^2 f(z)\mathrm{d}z\right] \tag{9.23}$$

在 Mammen 等(1999)给出的正则条件下(也可参考 Nielsen 和 Sperlich (2005)),可以证明

$$(nh_\alpha)^{1/2}[\hat{g}_\alpha(z_\alpha)-g_\alpha(z_\alpha)-h_\alpha^2\beta_\alpha(z_\alpha)] \xrightarrow{d} N(0,v_\alpha(z_\alpha)) \tag{9.24}$$

其中 β_α 在式(9.23)中定义而 $\beta(z)$定义为

$$\beta(z) = \kappa_2\sum_{s=1}^{q}\left[\frac{g_s^{(1)}(z_s)}{f(z)}\frac{\partial f(z)}{\partial z_s}+\frac{1}{2}g_s^{(2)}(z_s)\right]$$

且 $v_\alpha(z_\alpha) = \kappa\sigma_\alpha^2(z_\alpha)/f_\alpha(z_\alpha)$($\kappa_2 = \int k(v)v^2\mathrm{d}v, \kappa = \int k^2(v)\mathrm{d}v$)。

最后,可以用下式来估计 $g(z) = c_0 + \sum_{\alpha-1}^{q}g_\alpha(z_\alpha)$:

$$\hat{g}(z) = \bar{Y} + \sum_{\alpha=1}^{q}\hat{g}_\alpha(z_\alpha) \tag{9.25}$$

此外,Mammen 等(1999)证明了 $\hat{g}_\alpha(z_\alpha)$ 的不同组成部分是渐近相互独立的:因此,$\hat{g}(x)$ 的渐近分布易从式(9.24)中得到(见习题9.1)。Mammen 等也讨论了当 Z 具有紧支撑时边界修正核(boundary corrected kernels)的应用。

Mammen 等(1999)也考虑了使用 $g_\alpha(z_\alpha)$ 的一个局部线性估计量,使得目标函数(9.18)变为

$$\int\left[Y_i-\hat{c}_0-\sum_{\alpha=1}^{q}\hat{g}_\alpha(z_\alpha)-\sum_{\alpha=i}^{q}\hat{\theta}_\alpha(z_\alpha)(z_\alpha-Z_{\alpha i})\right]^2K_h(z,Z_i)\,dz \quad (9.26)$$

其中最小化是关于可加部分 $\hat{c}_0$、所有 $\hat{g}_\alpha(z_\alpha)$(满足$\int\hat{g}_\alpha(z_\alpha)\tilde{f}_\alpha(z_\alpha)dz_\alpha=0$),以及所有 $\hat{\theta}_\alpha(z_\alpha)$ 进行的,$\theta_\alpha(z_\alpha)$ 是 $g_\alpha(z_\alpha)$ 的一阶导数。

Mammen 等(1999)(也可参考 Mammen 和 Park(2005))证明了 $\hat{g}_\alpha$ 与 $\hat{\theta}_\alpha$ 满足下述等式: 293

$$\begin{aligned}\begin{bmatrix}\hat{g}_\alpha(z_\alpha)\\ \hat{\theta}_\alpha(z_\alpha)\end{bmatrix}&=-\begin{bmatrix}\hat{c}_0\\ 0\end{bmatrix}+\begin{bmatrix}\tilde{g}_\alpha(z_\alpha)\\ \tilde{\theta}_\alpha(z_\alpha)\end{bmatrix}\\ &=-\hat{M}_\alpha(z_\alpha)^{-1}\sum_{s\neq\alpha}^{q}\int\hat{S}_{s\alpha}(z_s,z_\alpha)\begin{bmatrix}\hat{g}_\alpha(z_\alpha)\\ \hat{\theta}_\alpha(z_\alpha)\end{bmatrix}dz_\alpha\end{aligned}$$

其中

$$\begin{aligned}\hat{c}_0&=\frac{1}{n}\sum_{i=1}^{q}Y_i-\sum_{\alpha=1}^{q}\int\hat{g}_\alpha(z_\alpha)\hat{f}_\alpha(z_\alpha)dz_\alpha\\ &\quad-\sum_{\alpha=1}^{q}\int\hat{\theta}_\alpha(z_\alpha)\hat{f}_\alpha^1(z_\alpha,z_\alpha)dz_\alpha\end{aligned} \quad (9.27)$$

$$\hat{M}_\alpha(z_\alpha)=n^{-1}\sum_{i=1}^{n}k_{h_\alpha}(z_a,Z_{\alpha i})\begin{bmatrix}1, & Z_{\alpha i}-z_\alpha\\ Z_{\alpha i}-z_\alpha, & (Z_{\alpha i}-z_\alpha)^2\end{bmatrix} \quad (9.28)$$

$$\begin{aligned}\hat{S}_{s\alpha}(z_s,z_\alpha)&=n^{-1}\sum_{i=1}^{n}k_{h_\alpha}(z_\alpha,Z_{si})k_{h_\alpha}(z_\alpha,z_{\alpha i})\\ &\quad\times\begin{bmatrix}1, & Z_{si}-z_s\\ Z_{\alpha i}-z_\alpha, & (z_s-Z_{si})(Z_{\alpha i}-z_\alpha)\end{bmatrix}\end{aligned} \quad (9.29)$$

$$\hat{f}_\alpha^1(z_\alpha)=n^{-1}\sum_{i=1}^{n}k_{h_\alpha}(z_s,Z_{si})(Z_{\alpha i}-z_\alpha),\text{and}$$

$$\hat{f}_\alpha(z_\alpha)=n^{-1}\sum_{i=1}^{n}k_{h_\alpha}(z_s,Z_{si})$$

另外，$\tilde{g}_\alpha(z_\alpha)$和$\tilde{\theta}_\alpha(z_\alpha)$是通过$Y_i$对$Z_{\alpha i}$回归得到的局部线性拟合；也就是说，它们最小化了目标函数

$$\sum_{i=1}^{n}[Y_i - \tilde{g}_\alpha(z_\alpha) - \tilde{\theta}_\alpha(z_\alpha)(Z_{\alpha i} - z_\alpha)]^2 k_{h_\alpha}(z_s, Z_{si}) \tag{9.30}$$

通过施加正则化条件

$$\int \hat{g}_\alpha(z_\alpha)\hat{f}_\alpha(z_\alpha)\mathrm{d}z_\alpha + \int \hat{\theta}_\alpha(z_\alpha)\hat{f}_\alpha^1(z_\alpha)\mathrm{d}z_\alpha = 0 \tag{9.31}$$

$\hat{c}_0, \hat{g}_1(z_1), \cdots, \hat{g}_q(z_q)$，以及$\hat{\theta}_1(z_1), \cdots, \hat{\theta}_q(z_q)$的定义可以变为唯一的。

294

平滑反向拟合估计是经由式(9.27)的迭代应用得到的，其中，当左边部分是第$[l+1]$步迭代时，右边部分是第$[l+1]$或者第$[l]$步，这取决于是$s<\alpha$还是$s>\alpha$。注意由式(9.31)得到$\hat{c}_0 = n^{-1}\sum_{i=1}^{n} Y_i$。

Mammen 等(1999)推导出了$\hat{g}_\alpha(z_\alpha)$的渐近分布。下面我们陈述一下 Nielsen 和 Sperlich(2005)给出的一个简单的版本。

$$\sqrt{nh_\alpha}\left[\hat{g}_\alpha(z_\alpha) - g_\alpha(z_\alpha) - v_\alpha - h_\alpha^2 \frac{\kappa^2}{2}\mu_\alpha(z_\alpha)\right] \xrightarrow{d} N(0, V_\alpha(z_\alpha)) \tag{9.32}$$

其中

$$v_\alpha = \int g_\alpha(z_\alpha) k_{h_\alpha}(z_\alpha, v) f_\alpha(v)\,\mathrm{d}v\mathrm{d}z_\alpha$$

$$\mu_\alpha(z_\alpha) = g_\alpha^{(2)}(z_\alpha) - \int g_\alpha^{(2)}(z_\alpha) f_\alpha(z_\alpha)\,\mathrm{d}z_\alpha$$

$$V_\alpha(z_\alpha) = \kappa\sigma_\alpha^2(z_\alpha)/f_\alpha(z_\alpha)$$

$$\kappa_2 = \int k(v)v^2\,\mathrm{d}v, \kappa = \int k^2(v)\,\mathrm{d}v$$

也可以通过矩阵形式来表示其联合分布，即

$$\begin{bmatrix} \sqrt{nh_1}\left[\hat{g}_1(z_1) - g_1(z_1) - v_1 - h_1^2 \frac{\kappa_2}{2}\mu_1(z_1)\right] \\ \vdots \\ \sqrt{nh_q}\left[\hat{g}_q(z_q) - g_q(z_q) - v_q - h_q^2 \frac{\kappa_2}{2}\mu_q(z_q)\right] \end{bmatrix} \xrightarrow{d} N(\mathbf{0}, \mathrm{diag}(V_\alpha(z_\alpha))) \tag{9.33}$$

其中$\mathbf{0} = (0, \cdots, 0)'$是$q \times 1$的零向量，$\mathrm{diag}(V_\alpha(z_\alpha))$是一$q \times q$的对角矩阵，其第$\alpha$个对角线元素等于$V_\alpha(z_\alpha)(\alpha = 1, \cdots, q)$，而非对角线元素均为零。

式(9.33)表明$\hat{g}_\alpha(z_\alpha)$的各个组成部分是渐近相互独立的。

Nielsen 和 Sperlich(2005)建议使用去一最小二乘交错鉴定法来选择平滑参数，而 Mammen 和 Park(2005)则推荐使用最小化惩罚残差平方和(penalized

sum of squared residuals)方法,即他们推荐选择 $h_1,\cdots,h_q$ 来最小化目标函数

$$\mathrm{PLS}(h) = \mathrm{RSS}(h)\left[1 + 2\sum_{\alpha=1}^{q}\frac{1}{nh_\alpha}k(0)\right] \tag{9.34}$$

其中

295

$$\mathrm{RSS}(h) = n^{-1}\sum_{i=1}^{n}\left[Y_i - \hat{c}_0 - \sum_{\alpha=1}^{q}\hat{g}_\alpha(z_{\alpha i})\right]^2 \tag{9.35}$$

本章我们并不讨论有交互项的可加模型的核估计,或者一般可加模型的导数估计。然而,读者可以在 Sperlich,Tjøstheim 和 Yang(2002)以及 Yang,Sperlich 和 Härdle(2003)中找到相关的讨论。另外,对于可加回归模型的有效和快速样条反向拟合核平滑(spline-backfitted kernel smoothing),可参考 Wang 和 Yang(2005)。

9.1.5 有连接函数的可加模型

一个更一般的可加函数是有一个已知连接函数的可加模型,如下:

$$Y_i = G\left[c_0 + \sum_{\alpha=1}^{q}g_\alpha(Z_{\alpha i})\right] + u_i \tag{9.36}$$

其中 $G(\cdot)$是一已知的连接函数。当 $G(\cdot)$为恒等函数时,式(9.36)退化为式(9.1)。在实际中,$G(\cdot)$可能是指数函数或对数函数等。

Linton 和 Härdle(1996)提出了一种边际积分方法来估计(9.36)。令 $m(z) = E(Y|z)$,$M(\cdot) = G^{-1}(\cdot)$;那么式(9.36)可以写成

$$M[m(z)] = c_0 + \sum_{\alpha=1}^{q}g_\alpha(z_\alpha) \tag{9.37}$$

定义

$$\phi_\alpha(z_\alpha) = \int M[m(z_\alpha, z_{\underline{\alpha}})]f_{\underline{\alpha}}(z_{\underline{\alpha}})\,\mathrm{d}z_{\underline{\alpha}} \tag{9.38}$$

其中 $z_{\underline{\alpha}} = (z_1,\cdots,z_{\alpha-1},z_{\alpha+1},\cdots,z_q)$,$f_{\underline{\alpha}}(z_{\underline{\alpha}})$为 $z_{\underline{\alpha}}$ 的概率密度函数。从式(9.38)我们知道 ϕ_α 与 g_α 的不同只相差一个可加的常数 c_0。Linton 和 Härdle(1996)建议基于以下多维局部常数核估计量来估计 ϕ_α:

$$\hat{m}(z_\alpha, z_{\underline{\alpha}}) = \frac{n^{-1}\sum_i Y_i k_{h_\alpha}(z_\alpha, Z_{\alpha i})W_{h_{\underline{\alpha}}}(z_{\underline{\alpha}}, Z_{\underline{\alpha} i})}{n^{-1}\sum_i k_{h_\alpha}(z_\alpha, Z_{\alpha i})W_{h_{\underline{\alpha}}}(z_{\underline{\alpha}}, Z_{\underline{\alpha} i})}, \tag{9.39}$$

其中 $k(\cdot)$为一个一元二阶核函数,且

$$W_{h_{\underline{\alpha}}}(v) = \prod\nolimits_{s\neq\alpha}^{q} h_s^{-1}w((z_s - Z_{si})/h_s)$$

其中 $w(\cdot)$是 $d > q-1$ 阶的核函数(当 $q > 2$ 时为高阶核)。这时,可以用式(9.38)的样本模拟来估计 $\phi_\alpha(z_\alpha)$,即

296

$$\tilde{\phi}_\alpha(z_\alpha) = \frac{1}{n}\sum_{i=1}^{n} M[\hat{m}(z_\alpha, Z_{\underline{\alpha}i})] f_{\underline{\alpha}}(Z_{\underline{\alpha}i}) \tag{9.40}$$

当 M 为恒等函数时,$\tilde{\phi}_\alpha(z_\alpha)$对 Y_i 是线性的。然而,一般地,$\tilde{\phi}_\alpha(z_\alpha)$是 Y_1 的非线性函数。常数项 c_0 的估计为

$$\tilde{c}_0 = \frac{1}{nq}\sum_{\alpha=1}^{q}\sum_{i=1}^{n}\tilde{\phi}_\alpha(Z_{\alpha i}) \tag{9.41}$$

且 $g_\alpha(z_\alpha)$由下式来估计:

$$\tilde{g}_\alpha(z_\alpha) = \tilde{\phi}_\alpha(z_\alpha) - \tilde{c}_0 \tag{9.42}$$

$m(z)$的最终估计由下式给出($G = M^{-1}$):

$$\hat{m}(z) = G\left[\sum_{\alpha=1}^{q}\tilde{g}_\alpha(z_\alpha) + \tilde{c}_0\right] \tag{9.43}$$

在某些正则条件下,Linton 和 Härdle(1996)证明了如下结果。

$$\sqrt{nh_\alpha}[\tilde{\phi}_\alpha - \phi_\alpha(z_\alpha) - h_\alpha^2\mu_\alpha(z_\alpha)] \xrightarrow{d} N(0, V_\alpha(z_\alpha)) \tag{9.44}$$

其中

$$\mu_\alpha(z_\alpha) = (\kappa_2/2)\{g_\alpha^{(2)}(z_\alpha)\int M^{(1)}[m(z)]f_{\underline{\alpha}}(z_{\underline{\alpha}})dz_{\underline{\alpha}}$$
$$+ 2g_\alpha^{(1)}(z_\alpha)\int M^{(1)}[m(z)]\frac{\partial \ln f(z)}{\partial z_\alpha}f_{\underline{\alpha}}(z_{\underline{\alpha}})dz_{\underline{\alpha}}$$

$$V_\alpha(z_\alpha) = \kappa\int (M^{(1)}[m(z)])^2\sigma^2 f_{\underline{\alpha}}^2(z_{\underline{\alpha}})\frac{1}{f_\alpha(z_\alpha)}dz_{\underline{\alpha}}$$

$V_\alpha(z_\alpha)$的一个一致估计量为

$$\tilde{V}_\alpha = \sum_{i=1}^{n}\tilde{\delta}_i^2\tilde{u}_i^2$$

297 其中

$$\tilde{\delta}_i = n^{-1}\sum_{i=1}^{n} M^{(1)}[\tilde{m}(z_\alpha, Z_{\underline{\alpha}_j})]w_i(z_\alpha, Z_{\underline{\alpha}i})$$

$$w_i(z_\alpha, Z_{\underline{\alpha}i}) = \frac{k_{h_\alpha}(z_\alpha - Z_{\alpha i})W_{h_{\underline{\alpha}}}(z_{\underline{\alpha}} - Z_{\underline{\alpha}i})}{\sum_{j=1}^{n}k_{h_\alpha}(z_\alpha - Z_{\alpha i})W_{\bar{h}}(z_{\underline{\alpha}}, Z_{\underline{\alpha}i})}$$

且 $\tilde{u}_i = Y_i - \tilde{m}(Z_i)$。

正如 Horowitz 和 Mammen(2004)观察到的那样,上述边际积分的基本估计量有一个不好的性质,即对 $w(\cdot)$必须应用一个更高阶的核,所取阶数与 q 有关(即 q 越大,$w(\cdot)$要求的阶数越高)。这是由于 ,在初始阶段,人们必须非参数估计一个 q 维函数 $m(z)$(即在初始阶段没有施加可加结构)。Horowitz 和 Mammen 建议初始期使用级数方法来估计 $m(z)$,然后再用局部线性核方法来估计

各个 $g_\alpha(\cdot)$ 函数,我们将在第 15 章讨论 Horowitz 和 Mammen 的方法并介绍非参数级数方法。Horowitz 和 Lee(2005)进一步把 Horowitz 和 Mammen 的结果扩展到了可加分位数回归模型上。

9.2 一个可加的部分线性模型

尽管上述介绍的可加模型有可以不用遭受维数诅咒的优点,但是它却是一个很有局限的模型,这是因为它不允许交互项的出现。一个可加部分线性模型可以避免这个问题且能保留"一维"非参数的收敛速率。考虑一个具有下列形式的模型:

$$Y_i = \beta_0 + X_i'\beta + g_1(Z_{1i}) + \cdots + g_q(Z_{qi}) + u_i \tag{9.45}$$

其中 X_i 是 $q\times 1$ 随机变量向量,$\beta=(\beta_1,\cdots,\beta_q)'$为 $q\times 1$ 未知参数向量,X_i 可以包含涉及$(Z_{1i},\cdots,Z_{qi})$的交互项,β_0 是一标量参数,所有的 $Z_{\alpha i}$都是一元连续变量,$g_\alpha(\cdot)(\alpha=1,\cdots,q)$为未知平滑函数。观测值$\{Y_i,X_i',Z_{1i},\cdots,Z_{qi}\}_{i=1}^n$都是独立同分布的。我们施加条件 $E[g_\alpha(Z_{\alpha i})]=0(\alpha=i,\cdots q)$来识别各个$g_1(\cdot),\cdots,g_q(\cdot)$。

298

定义 $v_i=X_i-E(X_i|Z_i)$,我们说 X 不是$(Z_i,\cdots,Z_q)$的确定性函数,若

$$E[v_iv_i'] \text{ 是正定的} \tag{9.46}$$

有了条件(9.46),就可以用第 6 章讨论的方法来得到 β 的一个$\sqrt{n}$—致估计量,记为 $\tilde{\beta}$。然后,人们通过把模型改写为以下形式就可以估计可加函数:

$$Y_i - X_i'\tilde{\beta} = c + \sum_{\alpha=1}^{q} g_\alpha(Z_{\alpha i}) + \varepsilon_i$$

其中 $\varepsilon_i=u_i+X_i'(\beta-\tilde{\beta})$。由于 $\tilde{\beta}-\beta=O_p(n^{-1/2})$比非参数可加函数估计量的收敛速度更快,因此当$\beta$已知时对应的 $g_\alpha(z_\alpha)$的非参数估计量也具有同样的渐近性质。因此,它们具有和 9.1 节讨论的非参数可加模型(没有线性部分)同样的渐近分布。

我们正式总结这个估计量如下。定义 $v_i=X_i-E(X_i|Z_i)$。当 $E(v_iv_i')$正定时,可以用如下简单的二步估计法来估计这个模型:

(i) 首先,忽略掉模型的可加结构并用 Robinson(1988)的方法来估计β,即关于 $X_i-\hat{E}(X_i|Z_i)$对 $Y_i-\hat{E}(Y_i|Z_i)$进行回归得到β的一个半参数估计量,记为$\tilde{\beta}$,其中 $\hat{E}(Y_i|Z_i)$和 $\hat{E}(X_i|Z_i)$分别为 $E(Y_i|Z_i)$和 $E(X_i|Z_i)$的核估计量。

(ii) 其次,改写模型为 $Y_i-X_i'\tilde{\beta}=g_i(Z_{1i})+\cdots+g_{qi}(Z_{qi})+\varepsilon_i$,并如 9.1 节讨论的那样估计可加函数 $g_l(\cdot)$。

$\tilde{\beta}$ 的渐近分布在第 7 章中已经讨论过，而相应的 $g_\alpha(z_\alpha)$ 的估计量，记为 $\tilde{g}_\alpha(z_\alpha)$，其渐近分布和 9.1 节中所讨论的相同 。这是因为 $\tilde{\beta}-\beta=O_p(n^{-1/2})$，其收敛速度比非参数的收敛速度要快；因此用 β 来替换 $\tilde{\beta}$ 并不影响 $\hat{g}_\alpha(z_\alpha)$ 的渐近分布。

299 以上过程的优点是计算上比较简单。然而，这种方法也存在一个问题。当 $E(v_i v_i')$ 为非正定时，这个过程就不能用来估计 β。考虑 $q=2$ 的简单情形，模型具有以下形式：

$$Y_i=\beta_0+(Z_{1i}Z_{2i})\beta+g_1(Z_{1i})+g_2(Z_{2i})+u_i \tag{9.47}$$

其中我们有 $X_i=Z_{1i}Z_{2i}$。在这种情况下，X_i 是 Z_i 的确定性函数，且有 $v_i=X_i-E(X_i|Z_i)=X_i-X_i=0$。也即以上过程是不能应用于此情形的。

然而，条件(9.46)是一个很强的假定，它排除了 X_i 为 $(Z_{1i},\cdots,Z_{qi})$ 的确定性函数(但非可加的)的情形。例如，对于 $q=2$ 我们希望使 X_i 含有交互项如 $X_i=Z_{1i}Z_{2i}$，但这样条件(9.46)就不再成立。

Fan 等(1998)以及 Fan 和 Li(2003)提出了一个基于边际积分方法的 β 的 $\sqrt{n}$一致估计量。他们的方法具有不需要依赖条件(9.46)的优点。然而，如前所述，边际积分方法在计算上是非常冗余的。Schick(1996)以及 Manzan 和 Zerom(2005)提出使用一个计算上有效率的方法来估计 β。下面我们来描述 Manzan 和 Zerom 的方法。

9.2.1 一个简单的二步方法

对任意随机变量(向量) ξ_i，定义

$$\xi_{\alpha i}=E[\xi_i w_\alpha(Z_i)\mid Z_{\alpha i}]\equiv\int E[\xi_i\mid Z_i=z]f_{\underline{\alpha}}(z_{\underline{\alpha}})\mathrm{d}z_{\underline{\alpha}}$$

其中 $w_\alpha(z)=f_\alpha(z_\alpha)f_{\underline{\alpha}}(z_{\underline{\alpha}})/f(z_\alpha,z_{\underline{\alpha}})$，$\alpha=1,\cdots,q$，并定义 ξ_i 在可加函数空间上的投射为

$$E_A(\xi_i)\sum_{\alpha=1}^{q}\xi_{\alpha i} \tag{9.48}$$

将投射(9.48)应用于式(9.45)可得

$$E_A(Y_i)=E_A(X_i)'\beta+\sum_{s=1}^{q}g(Z_{si}) \tag{9.49}$$

用式(9.45)减去(9.49)，我们得到

$$Y_i-E_A(Y_i)=(X_i-E_A(X_i))'\beta+u_i \tag{9.50}$$

我们可以用最小二乘法来估计 β。然而，我们必须先得到 $E_A(Y_i)$ 和 $E_A(X_i)$ 的一致估计量。由 9.1 节给出的结果，我们知道它们可以用下式被一致估

300

计出：

$$\hat{E}_A(Y_i) = \sum_{\alpha=1}^{q} \hat{Y}_{\alpha i}, \hat{E}_A(X_i) = \sum_{\alpha=i}^{q} \hat{X}_{\alpha i} \tag{9.51}$$

其中

$$\hat{\xi}_{\alpha i} = \frac{1}{nh_\alpha} \sum_{j=1}^{n} \xi_j k\left(\frac{Z_{\alpha j} - Z_{\alpha i}}{h_\alpha}\right) \frac{\hat{f}_{\underline{\alpha}}(Z_{\alpha j})}{\hat{f}(Z_{\alpha j}, Z_{\underline{\alpha} j})}, \alpha = 1, \cdots, q \tag{9.52}$$

且由式(9.25)中 $\xi_i = Y_i$ 或 $\xi_i = X_i$，我们可分别得到 $\hat{E}_A(Y_i)$ 和 $\hat{E}_A(X_i)$。

因此，一个 β 的可行估计量为

$$\hat{\beta} = S^{-1}_{X-\hat{E}_A(X)} S_{X-\hat{E}_A(X), Y-\hat{E}_A(Y)} \tag{9.53}$$

其中 $S_{C,D} = n^{-1} \sum_{i=1}^{n} C_i D_i', S_C = S_{C,C}$。Manzan 和 Zerom(2005)给出了 $\hat{\beta}$ 的渐近分布如下。

定理 9.2　在与 Manzan 和 Zerom(2005)给出的相同的正则条件下，有

$$\sqrt{n}(\hat{\beta} - \beta) \rightarrow N(0, \Sigma) \text{ 依分布收敛}$$

其中 $\Sigma = \Phi^{-1}\Omega\Phi^{-1}, \Phi = E[\eta_i\eta_i'], \eta_i = X_i - E_A(X_i), \Omega = E[u_i^2\eta_i\eta_i']$。

Σ 的一个一致估计量由 $\hat{\Sigma} = \hat{\Phi}^{-1}\hat{\Omega}\hat{\Phi}^{-1}$ 给出，其中

$$\hat{\Phi} = n^{-1} \sum_{i=1}^{n} [X_i - \hat{E}_A(X_i)][X_i - \hat{E}_A(X_i)]$$

$$\hat{\Omega} = n^{-1} \sum_{i=1}^{n} \hat{u}_i^2 [X_i - \hat{E}_A(X_i)][X_i - \hat{E}_A(X_i)]'$$

$$\hat{u}_i = Y_i - \hat{E}_A(Y_i) - (X_i - \hat{E}_A(X_i))'\hat{\beta}$$

当误差是条件同方差时，对这个模型，$\hat{\beta}$ 在其渐近方差达到半参数有效边界的意义下是半参数有效的(见 Chamberlain(1992))。

正定的 Φ 是 β 的一个识别条件。若 X_i 不是 $Z_{\alpha i}$ 的可加函数，它允许 X_i 为 $(Z_{1i}, \cdots, Z_{qi})$ 的确定性函数。更确切地说，考虑在式(9.47)中 $q = 2$ 且 $X_i = Z_{1i}Z_{2i}$ 的简单情形，如下：

$$Y_i = \beta_0 + (Z_{1i}Z_{2i})\beta + g_1(Z_{1i}) + g_2(Z_{2i}) + u_i \tag{9.54}$$

模型(9.54)是不受维数诅咒的影响的，这是由于它只涉及了一个一维的非参数函数 $g_\alpha(\cdot)(\alpha = 1, 2)$。同样，它比没有交互项的可加模型更一般化(即式(9.45)允许交互项以参数化形式进入模型)。　301

继续这个简单情形($q = 2$)，我们接下来估计各个非参数部分 $g_a(z_\alpha)$。给定 $\sqrt{n}$—致估计量 $\hat{\beta}$，我们可以重写式(9.45)为

$$Y_i - X_i'\hat{\beta} = \beta_0 + \sum_{\alpha=1}^{q} g_\alpha(Z_{\alpha i}) + \varepsilon_i \tag{9.55}$$

其中 $\varepsilon_i = u_i + X_i'(\beta - \hat{\beta})$。

式(9.55)实质上是一个以 $Y_i - X_i'\hat{\beta}$ 为新的因变量,以$[u_i + X_i'(\beta - \hat{\beta})]$为新的(复合)误差的可加回归模型。因此,人们可以通过对式(9.6)和式(9.7)讨论的部分线性模型进行边际积分来估计 $g_\alpha(z_\alpha)$,其中 Y_i 替换为 $Y_i - X_i'\hat{\beta}$。令 $\hat{g}_\alpha(z_\alpha)$表示得到的 $g_\alpha(z_\alpha)$的估计量。从定理 9.1 和 $\hat{\beta} - \beta = O_p(n^{-1/2})$的事实容易知道 $\hat{g}_\alpha(z_\alpha)$的渐近分布与定理 9.1 给出的 $\tilde{g}_\alpha(z_\alpha)$的渐近分布是相同的。

截距项 β_0 可以用 $\hat{\beta}_0 = \bar{Y} - \bar{X}_i'\hat{\beta}$ 来$\sqrt{n}$一致估计,其中 $\bar{Y} = n^{(-1)}\sum_{i=1}^{n} Y_i$,$\bar{X} = n^{-1}\sum_{i=1}^{n} X_i$。条件均值函数 $E[Y_i \mid X_i = x, Z_{1i} = z_i, \cdots, Z_{qi} = z_q]$ 可以用 $\hat{\beta}_0 + x'\hat{\beta} + \sum_{\alpha=1}^{q}\hat{g}_\alpha(z_\alpha)$ 来估计,误差项用 $\hat{u}_i = Y_i - \hat{\beta}_0 - X_i'\hat{\beta} - \sum_{\alpha=1}^{q}\hat{g}_\alpha(Z_{\alpha i})$ 来估计。

9.3 一个半参数变(平滑)系数模型

在前面我们讨论了一个部分线性模型,形式为

$$Y_i = \alpha(Z_i) + X_i'\beta_0 + u_i \tag{9.56}$$

其中 $\alpha(\cdot)$是一未知函数,β_0 是一 $r\times 1$ 未知参数向量。

本节中我们考虑一个更一般的半参数回归模型,所谓的半参数平滑系数模型,它嵌套了部分线性模型作为一种特殊情形。平滑系数模型形式为

$$Y_i = \alpha(Z_i) + X_i'\beta(Z_i) + u_i \tag{9.57}$$

其中 $\beta(z)$是关于 z 的未指定平滑函数的向量。当 $\beta(z) = \beta_0$ 时,式(9.57)退化为部分线性模型(9.56)。

例如,当为一个截面生产函数建模时,平滑系数模型是适合的,模型右端变量是劳动力、资本和厂商的 R&D 投入。如果我们令 $X_i = (\text{labor}_i, \text{capital}_i)$,$Z_i = \text{R\&D}_i$,那么式(9.57)意味着劳动力和资本投入的系数可能随着厂商的 R&D 投入而变化。于是劳动力与资本两者的边际生产均依赖厂商的 R&D 值,结果,规模报酬可能也是 R&D 的函数。部分线性模型(9.56)假定斜率系数 β 关于 R&D 是不变的,于是 R&D 变量只能影响生产前沿水平。这种情况下,R&D 被称作对生产前沿具有"中性"效应。与式(9.56)不同,平滑系数模型 (9.57) 允许 R&D"非中性"地影响随机前沿。

9.3.1 平滑系数函数的一个局部常数估计量

为了避免引入过多的符号,我们稍微改变一下符号定义,将式(9.57)表达

为更紧致的形式

$$Y_i = X_i'\beta(Z_i) + u_i \tag{9.58}$$

其中 X_i 为一 $p\times 1$ 向量,可能还包含一个常数作为其第一个元素(这样就与式(9.57)等价),$\beta(z)$ 为 z 的一个 $p\times 1$ 维函数,Z_i 的维数为 q。

对式(9.58)两边同乘以 X_i 并取期望 $E(\cdot\mid Z_i)$ 得到 $E[X_iY_i\mid Z_i]=E[X_iX_i']\beta(Z_i)$,有

$$\beta(z) = [E(X_iX_i'\mid z)]^{-1}E[X_iY_i\mid z] \tag{9.59}$$

式(9.59)意味着如下的 $\beta(z)$ 的局部常数最小二乘估计量:

$$\hat{\beta}(z)\Big[\sum_{j=1}^{n} X_jX_j'K\Big(\frac{Z_j-z}{h}\Big)\Big]^{-1}\sum_{j=1}^{n} X_jY_jK\Big(\frac{Z_j-z}{h}\Big) \tag{9.60}$$

上述局部最小二乘估计量背后的直觉应该是显而易见的。在 z 为一标量,$K(\cdot)$ 为均匀核的情形下,式(9.60)变为

303

$$\hat{\beta}(z)[\sum_{|Z_j-z|\leqslant h} X_jX_j']^{-1}\sum_{|Z_j-z|\leqslant h} X_jY_j \tag{9.61}$$

这里,$\hat{\beta}(z)$ 只是一个 Y_j 对 X_j 回归的最小二乘回归量,使用的观察集为 (X_j,Y_j),其相应的 Z_j 接近 $z(|Z_j-z|\leqslant h)$。由于 $\beta(z)$ 是 z 的一个平滑函数,当 $|Z_j-z|$ 比较小时,$|\beta(Z_j)-\beta(z)|$ 也比较小。当 $\beta(Z_j)$ 很接近 $\beta(z)$ 时,nh 很大这个条件保证了我们有足够数量的观测值落在区间 $|Z_j-z|\leqslant h$ 中。因此,在 $h\to 0, nh\to\infty$ ($Z_i\in\mathbb{R}^q$ 时,$nh_1\cdots h_q\to\infty$)这样的条件下,我们能够证明局部最小二乘估计量 $\hat{\beta}$ 是 $\beta(z)$ 的一个一致估计量。

下述定理建立了 $\hat{\beta}(z)$ 的一致性和渐近正态性。

定理 9.3 在一定的正则条件(见 Li, Huang, Li 和 Fu(2002))下,对 z 的一固定值且满足 $f_z(z)>0$($f_z(\cdot)$ 为 Z_i 的边缘概率密度函数),我们有

$$\sqrt{nh_1\cdots h_q}\Big[\hat{\beta}(z)-\beta(z)-\sum_{s=1}^{q}h_s^2B_s(z)\Big]\to N(0,\Omega_z) \text{ 依分布收敛}$$

其中假定 $M(z)\overset{\text{def}}{=}f_z(z)E[X_iX_i'\mid Z_i=z]$ 是正定的,$B_s(z)=\kappa_2M(z)^{-1}E[X_iX_i'\{\beta_s(z)f_s(X_i,Z_i)/f(X_i\mid Z_i=z)+(1/2)f_z(Z_i)\beta_{ss}(Z_i)\}\mid z]$,$\kappa_2=\int k(v)v^2\mathrm{d}v$,$\beta_s(z)=\partial\beta(z)/\partial z_s$,$\beta_{ss}(z)=\partial^2\beta(z)/\partial z_s^2$,$\Omega_z=M(z)^{-1}V_zM(z)^{-1}$,$V_z=\kappa^qf_z(z)E[X_iX_i'\sigma^2(X_i,Z_i)\mid Z_i=z]$,$\sigma^2(X_i,Z_i)=E(u_i^2\mid X_i,Z_i)$。$\Omega_z$ 的一个一致估计量在习题 9.2 节中给出。

定理 9.3 的证明在 9.3.5 节中给出。

9.3.2 平滑系数函数的一个局部线性估计量

下述讨论中的模型与式(9.58)给出的模型是相同的。为了表述简洁,我们

假定 Z_i 为一标量，于是 $\beta_s(z)$ 的局部线性估计量为 $\hat{\beta}_s(z) = \hat{a}_s(z)$，其中 $\{(\hat{a}_s(z),\hat{b}_s(z))\}_{s=1}^{p}$ 最小化了下述加权平方和：

$$\sum_{i=1}^{n}[Y_i - \sum_{s=1}^{p}\{a_s + b_s(Z_i - z)\}X_{is}]^2 K_h(Z_i,z) \tag{9.62}$$

304 其中 $k_h(\cdot) = h^{-1}K(\cdot/h)$。

令 $\delta(z) = (a_1(z),\cdots,a_q(z),b_1(z),\cdots,b_p(z))'$，又令 $\mathcal{X}$ 表示一 $n\times 2p$ 矩阵，其第 i 行为 $(X_i', X_i'(Z_i - z))$，再令 $\mathcal{K}$ 为一对角矩阵（维数为 n），其第 i 个对角线上元素为 $K_h(Z_i - z)$，同时 $\mathcal{Y} = (Y_1,\cdots,Y_n)'$，那么，由式(9.62)，根据最小二乘理论，$\delta(z)$ 的估计量为

$$\hat{\delta}(z) = (\mathcal{X}'\mathcal{K}\mathcal{X})^{-1}\mathcal{X}\mathcal{K}\mathcal{Y} \tag{9.63}$$

注意 $\hat{\delta}(z) = (\hat{a}_1(z),\cdots,\hat{a}_p(z),\hat{b}_1(z),\cdots,\hat{b}_p(z))'$，其中 $\hat{a}_s(z)$ 估计的是 $\beta_s(z)$，$\hat{b}_s(z)$ 估计的是 $\partial\beta(z)/\partial z_s$。

为了导出局部线性估计量 $\hat{\delta}(z)$ 的渐近分布，我们首先引入一些符号，令

$$\boldsymbol{S}_n = \boldsymbol{S}_n(z)\begin{bmatrix} S_{n,0}, & S_{n,1} \\ S_{n,1}, & S_{n,2}\end{bmatrix}, \boldsymbol{T}_n = \boldsymbol{T}_n(z) = \begin{bmatrix} T_{n,0} \\ T_{n,1}\end{bmatrix}$$

其中，对 $j = 1,2$，

$$S_{n,j} = S_{n,j}(z) = n^{-1}\sum_{i=1}^{n}X_iX_i'((Z_i - z)/h)^j K_h(Z_i,z)$$

$$T_{n,j} = T_{n,j}(z) = n^{-1}\sum_{i=1}^{n}X_iY_i((Z_j - z)/h)^j K_h(Z_i,z)$$

另外，令

$$\delta(z) = (a_1,\cdots,a_p,b_1,\cdots,b_p)'$$

$$\hat{\delta} = \hat{\delta}(z) = (\hat{a}_1(z),\cdots,\hat{a}_p(z),\hat{b}_1(z),\cdots,\hat{b}_p(z))^{\mathrm{T}}$$

使用这些符号我们可以将 $\hat{\delta}(z)$ 写作

$$\hat{\delta}(z) = \boldsymbol{H}^{-1}\boldsymbol{S}_n^{-1}\boldsymbol{T}_n \tag{9.64}$$

其中 $\boldsymbol{H} = \mathrm{diag}\{1,\cdots,1,h,\cdots,h\}$，它的 p 个对角线元素为 1，另外 p 个对角线元素为 h。

为了符号表示方便，对非负的整数 j，我们定义

$$\mu_j = \int_{-\infty}^{\infty} v^j k(v)\,\mathrm{d}v, \quad \nu_j = \int_{-\infty}^{\infty} v^j k^2(v)\,\mathrm{d}v$$

此外

$$\boldsymbol{\Omega}(z) = (w_{l,m})_{p\times p} = E[\mathcal{X}_i\mathcal{X}_i' \mid Z_i = z]$$

305 我们将假定核函数 $k(\cdot)$ 是对称的，于是 $\mu_1 = \int uk(u)\,\mathrm{d}u = 0$，同时我们也

假定 $\mu_0 = \int k(u)\mathrm{d}u = 1$，那么对于每个 $0 \leqslant j \leqslant 3$ 和 $1 \leqslant l, m \leqslant p$，可以证明(见习题9.4)

$$E[S_{n,j}(z)] \to f(z)\Omega(z)\mu_j,\ nh\mathrm{var}(S_{n,j}(z)_{l,m}) \to f(z)v_{2j}w_{l,m} \tag{9.65}$$

其中 $f(\cdot)$ 是 Z_i 的概率密度函数。式(9.65)的一个结果是

$$S_n \xrightarrow{p} f(z)S,\ S_{n,3} \xrightarrow{p} \mu_3 f(z)\Omega(z) \tag{9.66}$$

其中

$$S = \begin{bmatrix} \Omega(z), & 0 \\ 0, & \mu_2\Omega(z) \end{bmatrix}$$

Cai,Fan 和 Yao(2000)证明了如下结果：

定理 9.4 在一定的平滑和矩正则条件下，另假定 $f(z) > 0$，那么有($\hat{a}(z) = (\hat{a}_i(z), \cdots, \hat{a}_p(z))'$)

$$\sqrt{nh}\left[\hat{a}(z) - a(z) - \frac{h^2}{2}\mu_2 a^{(2)}(z)\right] \to N(0, \Theta(z)) \text{ 依分布收敛}$$

其中 $a^{(2)}(z)$ 是个 $p \times 1$ 向量，其第 s 项为 $a_s^{(2)}(z) = \partial^2 a_s(z)/\partial z^2$，其中

$$\Theta(z) = v_0\Omega^{-1}(z)\Omega^*(z)\Omega^{-1}(z)/f(z)$$

$$\Omega^*(z) = E[X_iX_i'\sigma^2(X_i, Z_i) \mid Z_i = z]$$

定理 9.4 意味着

$$\sqrt{nh}\left[\hat{a}_s(z) - a_s(z) - \frac{h^2}{2}\mu_2 a_s^{(2)}(z)\right] \to N(0, V_{s,0}) \text{ 依分布收敛}$$

其中 $V_{s,0} = v_0 e_{s,p}'\Omega^{-1}(z)\Omega^*(z)\Omega(z)e_{s,p}/f(z)$，$e_{s,p}$ 是一 $p \times 1$ 向量，其第 s 项为 1，其余为 0。

Xue 和 Yang(2006)给出了一个替代的方法来估计半参数变系数模型(如多项式样条)。

9.3.3 参数平滑系数模型的检验

306

当 $\beta(z) = \beta_0(z)$ 具有已知的函数形式时，可以用参数方法来估计这个模型。例如，若 $\beta(z) = \beta_0$ 为参数常数向量，我们可以用普通最小二乘法来估计模型。若 $\beta(z) = \beta_0(z, \gamma)$，其中 $\beta_0(\cdot)$ 具有已知的函数形式，γ 为一有限维参数，如果模型对参数是非线性的，我们可以用非线性最小二乘方法来估计模型。下面我们讨论怎样检验几乎处处 $\beta(z) = \beta_0(z, \gamma)$(对 z)。

1. 一个基于局部常数估计量的检验

令 $\hat{\gamma}$ 为 γ 的一个 $\sqrt{n}$ 一致估计量且 $\hat{\beta}_0(z) = \beta_0(z, \hat{\gamma})$，可以构造一个基于 $\hat{\beta}(z)$ 和 $\hat{\beta}_0(z)$ 的差的加权平方积分的检验统计量：

$$I_n = \int \{ [\hat{\beta}(z) - \hat{\beta}_0(z)]' A_n [\hat{\beta}(z) - \hat{\beta}_0(z)] \} \mathrm{d}z$$

其中 A_n 是一个正定矩阵。$\hat{\beta}(z)$ 是一个局部常数估计量，它有一个随机分母 $D_n(z)^{-1}$，这将使对 I_n 的渐近分析复杂化。Li 等(2002)建议选择 $A_n = D_n(z)'D_n(z)$ 来移除 I_n 中的随机分母。通过这样选择 A_n 我们有

$$\begin{aligned} I_n &= \int \{ D_n(z)[\hat{\beta}(z) - \hat{\beta}_0(z)] \}' D_n(z)[\hat{\beta}(z) - \hat{\beta}_0(z)] \mathrm{d}z \\ &= \frac{1}{n^2} \sum_i \sum_j \int X_i' [Y_i - X_i'\hat{\beta}_0(z)] X_j [y_j - X_j'\hat{\beta}_0(z)] K_{h,z_i,z} K_{h,z_j,z} \mathrm{d}z \\ &= \frac{1}{n^2} \sum_i \sum_j X_i' X_j \hat{u}_i \hat{u}_j \bar{K}_{h,z_i,z_j} \end{aligned}$$

其中 $\hat{u}_i = Y_i - X_i'\hat{\beta}_0(z)$，$\bar{K}_{h,z_i,z_j} = \prod_{s=1}^{q} h_s^{-1} \bar{k}((Z_{is} - Z_{js})/h_s)$，而 $\bar{k}(v) = \int k(u)k(u-v)\mathrm{d}u$ 是从 $k(\cdot)$ 推导出的二重卷积核。

事实上并没有必要去计算卷积核，这是由于可以简单地用一个标准二阶核 $K(\cdot)$ 来替代 $\bar{K}(\cdot)$。另外，也可以剔除 I_n 中 $i=j$ 的项，这样检验统计量在原假定下将渐近以零为中心。因此，Li 等(2002)给出的最终的检验估计量为

307

$$\hat{I}_n = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} X_i' X_j \hat{u}_i \hat{u}_j K_{h,z_i,z_j}$$

Li 等(2002)进一步证明了在原假定下，

$$\hat{J}_n = n(h_1 \cdots h_q)^{1/2} \hat{I}_n \Big/ \sqrt{\hat{\sigma}_0^2} \xrightarrow{d} N(0,1)$$

其中[①]

$$\hat{\sigma}_0^2 = \frac{2h_1 \cdots h_q}{n^2} \sum_{i=1}^{n} \sum_{j \neq i}^{n} (X_i' X_j)^2 \hat{u}_i^2 \hat{u}_j^2 K_{h,zi,zj}^2$$

也可以通过原始自举法来逼近 $\hat{J}_n$ 原假定下的分布。计算自举检验统计量的准确步骤留作一个习题。

2. 一个基于局部线性估计量的检验

Cai, Fan 和 Yao(2000)建议基于参数与半参数残差平方和的比较来检验参数变系数函数。定义

$$\mathrm{RSS}_{\mathrm{para}} = n^{-1} \sum_{i=1}^{n} [Y_i - \beta_1(Z_i, \hat{\gamma}) X_{i1} - \cdots - \beta_p(Z_i, \hat{\gamma}) X_{ip}]^2$$

$$\mathrm{RSS}_{\mathrm{semi}} = n^{-1} \sum_{i=1}^{n} [Y_i - \hat{\beta}_1(Z_i) X_{i1} - \cdots - \hat{\beta}_p(Z_i) X_{ip}]^2$$

---

① 这里的表达式与 Li 等(2002)中的完全相同，注意 $K_{h,ij} = (h_1 \cdots h_q)^{-1} K((z_j - z_i)/h)$。

其中 $\hat{\beta}_s(z)=\hat{a}_s(z)$ 是在9.3.2节中讨论的 $\beta_s(z)$ 的局部线性估计量。Cai, Fan和Yao提出使用以下检验统计量来检验几乎处处 $\beta(z)=\beta_0(z,\gamma)$ 的原假定：

$$TS_n=[RSS_{para}-RSS_{semi}]/RSS_{semi}=RSS_{para}/RSS_{semi}-1$$

在实际中，TS_n 值较大时拒绝原假定。Cai, Fan和Yao提出使用如下的参数自举过程来评测 TS_n 检验的 P 值。

(i) 从中心残差 $\{\hat{u}_i-n^{-1}\sum_{j=1}^{n}\hat{u}_j\}_{i=1}^{n}$ 的经验分布生成自举残差 u_i^*，其中

308

$$\hat{u}_i=Y_i-\hat{\beta}_1(Z_i)X_{i1}-\cdots-\hat{\beta}_p(Z_i)X_{ip}$$

定义

$$Y_i^*=\beta_1(Z_i,\hat{\gamma})X_{i1}+\cdots+\hat{\beta}_p(Z_i,\hat{\gamma})X_{ip}+u_i^*$$

(ii) 基于自举样本 $\{Z_i,X_i',Y_i^*\}_{i=1}^{n}$ 来计算自举法统计量 TS_n^*，重复步骤(i)与步骤(ii)非常多次，例如 $B=399$。

(iii) 当 TS_n 大于自举统计量 $\{TS_{n,j}^*\}_{j=1}^{B}$ 的 α 上分位点时，拒绝原假定。

也可以在步骤(i)中使用原始自举方法的两点分布来生成上面的 u_i^*，即 $u_i^*=[(1-\sqrt{5})/2]\hat{u}_i$ 的概率为 $r=(1+\sqrt{5})/(2\sqrt{5})$，$u_i^*=[(1+\sqrt{5})/2]\hat{u}_i$ 的概率为 $1-r$，于是得到的自举统计量对于条件异方差的出现是稳健的。

我们推荐感兴趣的读者参考Yang, Park, Xue和Härdle(2006)，它提供了一个基于边际积分方法针对可加模型中变系数结构的替代检验过程。Fan和Zhang(2000)推导出了变系数模型的同步置信区间，这也可以用于检验假定。

9.3.4　部分线性平滑系数模型

我们可以将前述模型推广到部分线性框架，考虑以下形式的模型：

$$Y_i-W_i'\gamma+X_i'\beta(Z_i)+u_i \tag{9.67}$$

其中 γ 是一个未知(常数)参数向量，而 $\beta(z_i)$ 是平滑系数函数。

若 γ 已知，我们能改写式(9.67)为

$$Y_i-W_i'\gamma=X_i'\beta(Z_i)+u_i \tag{9.68}$$

且像前面讨论的那样将 $Y_i-W'_i\gamma$ 作为新的因变量来估计 $\beta(Z_i)$，即

309

$$\hat{\beta}(Z_i)=\left[\sum_j X_jX_j'K_{h,z_j,z}\right]^{-1}\sum_j X_j[Y_j-W_j'\gamma]K_{h,z_j,z}$$

$$\equiv\hat{\beta}_y(Z_i)-\hat{\beta}_w(Z_i)'\gamma \tag{9.69}$$

其中

$$\hat{\beta}_y(Z_i)=\left[\sum_j X_jX_j'K_{h,z_j,z_i}\right]^{-1}\sum_j X_jY_jK_{h,z_j,z_i}$$

$$\hat{\beta}_w(Z_i)=\left[\sum_j X_jX_j'K_{h,z_j,z_i}\right]^{-1}\sum_j X_jW_jK_{h,z_j,z_i}$$

用式(9.69)中给出的估计量$\hat{\beta}(z_i)$来替代式(9.68)中的$\beta(z_i)$，我们得到

$$Y_i - X_i'\hat{\beta}_y(Z_i) = (W_i' - X_i'\hat{\beta}_w(Z_i)')\gamma + \text{error} \tag{9.70}$$

我们现在可以通过普通最小二乘方法来估计γ，通过$Y_i - X_i'\hat{\beta}_y(Z_i)$对$(W_i' - X_i'\hat{\beta}_w(Z_i)')$进行回归。令$\hat{\gamma}$表示得到的$\gamma$的(半参数)估计量，Fan 和 Huang(2005)证明

$$\sqrt{n}(\hat{\gamma} - \gamma) \to N(0, V) \text{ 依分布收敛} \tag{9.71}$$

其中$V = A^{-1}BA^{-1}$，其中$A = E[\zeta_i\zeta_i']$，$B = E[\sigma^2(X_i, Z_i, W_i)\zeta_i\zeta_i']$，$\zeta_i = W_i - E_{sm}(W_i)$，其中$E_{sm}(W_i)$定义为$W_i$在平滑系数函数空间上的投射，即$E_{sm}(W_i) = X_i'\beta_w(Z_i)$，而$\beta_w(\cdot) = \inf_{\beta(\cdot)} E\{[W_i - X_i'\beta(Z_i)]'[W_i - X_i'\beta(Z_i)]\}$。上述方法被称为“剖面似然”(profile likelihood)法。习题 9.3 给出了$\beta_w(Z_i)$的一个一致估计量，对读者来说应该比较容易构造一个V的一致估计量，这可以通过以样本均值$n^{-1}\sum_{i=1}^{n}(\cdot)$替代总体均值$E(\cdot)$来实现。

当误差项是条件同方差时，即$\sigma_u^2 = E(u_i^2 \mid X_i, Z_i, W_i) = E(u_i^2)$，$V$简化为$V = \sigma_u^2 A^{-1}$，这与$\gamma$的(正则)估计量的渐近方差下界相同。于是，当误差项是条件同方差时，$\hat{\gamma}$是γ的一个半参数有效估计量。

310

9.3.5 定理 9.3 的证明

证明：

$$\begin{aligned}
\hat{\beta}(z) &= \Big[\sum_j X_jX_j'K_{z_j,z}\Big]^{-1}\sum_j X_jY_jK_{z_j,z} \\
&= \Big[\sum_j X_jX_j'K_{z_j,z}\Big]^{-1}\sum_j X_j[X_j'\beta(Z_j) + u_j]K_{z_j,z} \\
&= \Big[\sum_j X_jX_j'K_{z_j,z}\Big]^{-1}\sum_j X_j[X_j'\beta(z) + X_j'(\beta(Z_j) - \beta(z)) + u_j]K_{z_j,z} \\
&= \beta(z) + \Big[\sum_j X_jX_j'K_{z_j,z}\Big]^{-1}\sum_j X_j[X_j'(\beta(Z_j) - \beta(z)) + u_j]K_{z_j,z} \\
&= \beta(z) + [D_n(z)]^{-1}\{A_{1n}(z) + A_{2n}(z)\}
\end{aligned}$$

其中

$$D_n(z) = n^{-1}\sum_j X_jX_j'K_{h,z_j,z}$$

$$A_{1n}(z) = n^{-1}\sum_j X_jX_j'(\beta(Z_j) - \beta(z))K_{h,z_j,z}$$

$$A_{2n}(z) = n^{-1}\sum_j X_ju_jK_{h,z_j,z}$$

下述结论(i)—(iii)可由定理9.3推出。

(i) $D_n(z) = n^{-1}\sum_j X_j X_j' K_{h,z_j,z} \xrightarrow{p} M_2$

(ii) $A_{1n}(z) = M_z \sum_{S=1}^{q} h_s^2 B_s(z) + o_p(\sum_{s=1}^{q} h_s^2)$

(iii) $\sqrt{nh_1\cdots h_q}A_{2n}(z) \xrightarrow{d} N(0, V_z)$

我们接下来证明这些结果。记 $H_q = h_1\cdots h_q, H_q^{-1} = (h_1\cdots h_q)^{-1}, h^2 = \sum_{s=1}^{q} h_s^2$，我们有　311

$$\begin{aligned}
\text{(i) } E(D_n(z)) &= H_q^{-1}E[X_iX_i'K_{z_i,z}] \\
&= H_q^{-1}\iint x_i x'_1 K((z_1 - z)/h)f(x_1, z_1)\mathrm{d}x_1\mathrm{d}z_1 \\
&= \iint x_i x_1' K(v)f(x_1, z + hv)\mathrm{d}v\mathrm{d}x_1 \\
&= \left[\int x_1x_1'f(x_1, z)\mathrm{d}x_1\right]\left[\int K(v)\mathrm{d}v + O(h^2)\right] \\
&= f_z(z)\left[\int x_1x_1'f(x_1 \mid z_1 = z)\mathrm{d}x_1\right][1 + O(h^2)] \\
&= f_z(z)E[X_iX_i' \mid Z_i = z] + o(1) = M_z + o(1)
\end{aligned}$$

类似地,我们可以证明 $\mathrm{var}(D_n(z)) = O((nh_1\cdots h_q)^{-1}) = o(1)$。于是,我们证明了

$$D_n(z) = M_z + o_p(1) \tag{9.72}$$

$$\begin{aligned}
\text{(ii) } E[A_{1n}(z)] &= E[X_iX_i'(\beta(Z_i) - \beta(z))K_{h,z_i,z}] \\
&= E\{E[X_iX_i' \mid Z_i](\beta(Z_i) - \beta(z))K_{h,z_i,z}\} \\
&= \int E[X_iX_i' \mid z_i]f(z_i)(\beta(z_i) - \beta(z))K_{h,z_i,z}\mathrm{d}z_i \\
&= \int M(z + hv)(\beta(z + hv) - \beta(z))K(v)\mathrm{d}v \\
&= \kappa_2\sum_{s=1}^{q} h_s^2[M_s(z)\beta_s(z) + (1/2)M(z)\beta_{ss}(z)] + O(\sum_{s=1}^{q} h_s^3) \\
&= \kappa_2 M(z)\sum_{s=1}^{q} h_s^2 B_s(z) + O(\sum_{s=1}^{q} h_s^3)
\end{aligned}$$

其中 $B_s(z)$ 按定理9.3中的定义。类似地,可以证明

$$\mathrm{var}(A_{1n}(z)) = O(\sum_{s=1}^{q} h_s^2(nh_1\cdots h_q)^{-1})$$

于是有

$$\left(A_{1n}(z) - M_2 \sum_{s=1} h_s^2 B_s(z)\right) = o_p\left(\sum_{s=1}^{q} h_s^2 + (nh_1 \cdots h_p)^{-1/2}\right) \qquad (9.73)$$

312　(iii) $\sqrt{nh_1 \cdots h_p} A_{2n}(z)$ 均值为零且其方差为

$$nH_q n^{-2} \sum_{i=1}^{n} E[X_i X_i' u_i^2 K_{h,z_i,z}^2]$$

$$= H_q \int f(x_1, z_1) \sigma_u^2(x_1, z_1) x_1 x'_1 K_{h,z_1,z}^2 \mathrm{d}x_1 \mathrm{d}z_1$$

$$= \left[\int x_1 x'_1 \sigma_u^2(x_1, z) f(x_1, z) \mathrm{d}x_1\right]\left[\int K^2(v) \mathrm{d}v\right] + O\left[\sum_{s=1}^{q} h_s\right]$$

$$= \kappa^q f_z(z) E[X_i X_i' \sigma_u^2(X_i, Z_i) \mid Z_i = z] + o(1) = V_z + o(1)$$

其中 $\kappa^q = \int K^2(v) \mathrm{d}v, H_q = h_1 \cdots h_q$ 。

可以直接验证 Liapunov 中心极限定理成立的条件满足。于是，

$$\sqrt{nH_q} A_{2n}(z) \xrightarrow{d} N(0, V_z) \qquad (9.74)$$

综合式(9.73)，式(9.72)与式(9.74)的结果，我们证明了

$$\sqrt{nH_q}\left[\hat{\beta}(z) - \beta(z) - \sum_{s=1}^{q} h_s^2 B_s(z)\right]$$

$$= [D_n(z)]^{-1} \sqrt{nH_q} \left\{A_{1n}(z) - M_z \sum_{s=1}^{q} h_s^2 B_s(z) + A_{2n}(z)\right\}$$

$$= [M_z + o_p(1)]^{-1} \{\sqrt{nH_q} A_{2n}(z) + o_p(1)\}$$

$$\xrightarrow{d} M_z^{-1} N(0, V_z) = N(0, M_z^{-1} V_z M_z^{-1})$$

9.4 习　题

习题 9.1　令 $\hat{g}(z)$ 如式(9.25)中所定义，假定 $h_1 = \cdots = h_q$，证明

$$\sqrt{nh}\left[\hat{g}(z) - g(z) - \frac{\kappa_2}{2} h^2 \sum_{s=1}^{q} g_{ss}(z_s)\right] \xrightarrow{d} N\left[0, \sum_{s=1}^{q} v_s(z_s)\right]$$

提示：由 $\hat{c}_0 - c_0 = O_p(n^{-1/2}) = o_p((nh)^{-1/2})$，可以用 c_0 替代 $\hat{c}_0$ 而不影响 $\hat{g}(z)$ 的渐近分布。另外，用到了当 $\alpha \neq s$ 时，$\hat{g}_\alpha(z_\alpha)$ 与 $\hat{g}_s(z_s)$ 渐近相互独立的事实。

313　**习题 9.2**　证明 $\hat{\Omega}_z = \hat{M}_z^{-1} \hat{V}_z \hat{M}_z^{-1}$ 是定理 9.3 中定义的 Ω_z 的一致估计量，其中 $\hat{M}_z = n^{-1} \sum_i X_i X_i' K_{h,z_i,z}$，有

$$K_{h,z_i,z} \equiv \prod_{s=1}^{q} h_s^{-1} k\left[\frac{Z_{is} - z_s}{h_s}\right]$$

$$\hat{V}_z = (h_1 \cdots h_q / n) \sum_{i=1}^{n} X_i X_i' \hat{u}_i^2 K_{h,z_i,z}^2$$

$$\hat{u}_i = Y_i - X_i' \hat{\beta}(Z_i)$$

习题 9.3　构造式(9.71)中所定义的渐近方差 V 的一个一致估计量。

提示:不需要对结果进行证明。构造 A 和 B 的一致估计量,分别记为 $\hat{A}$ 与 $\hat{B}$,这时有 $\hat{V} = \hat{A}^{-1}\hat{B}\hat{A}^{-1}$。注意 $\beta_w(z_i)$ 可以通过以下半参数变系数模型来一致估计

$$W_i = X_i' \beta_w(Z_i) + \text{error}$$

于是 $\hat{\beta}_w(Z_i)[\sum_{j=1}^{n} X_j X_j' K_{h,z_i,z_j}]^{-1} \sum_{j=1}^{n} X_j W_j K_{h,z_j,z_j}$。

习题 9.4　证明式(9.65)的第一部分,即证明对于 $j=0,1,2$,有

$$E[\boldsymbol{S}_{n,j}(z)] \to f(z)\Omega(z)\mu_j$$

第 10 章 选择模型

315

样本选择问题会经常出现在应用研究过程中。例如,在医药领域,那些选择参加临床试验的人可能比从一般总体中选取的人更有可能(或更不可能)对一个治疗产生反应,因为有过去治疗的原因。在社会科学中,类似的情况也有发生,例如,与随机选择的个体相比,选择参加职业培训项目的那些人更有可能(或更不可能)从项目中受益。在每一种情形中,人们可能会担心那些“选择进入”一个项目的人和从一般总体中随机选取的人相比,治疗(训练)的影响会有所不同。因此使用非代表性样本进行分析会使测量的处理效果变得混淆。或者研究者可能对社会经济特征影响所关注的某个结果的程度感兴趣,再次担心基于子样本的这些特征的影响是否对一般总体成立。

处理样本选择问题的具有开拓性的参数方法可参见 Heckman (1976,1979),他提出了一个被广泛运用的方法,其中描述了结果的过程且被实施,在第二阶段使用来自于此的信息得到相关参数的一致估计。有很多处理样本选择的精巧方法被提了出来,本章我们考虑一些用于处理这种问题的半参数和非参数方法。

316

10.1 半参数类型 2 Tobit 模型

考虑下列潜变量方程:

$$Y_{1i}^* = X_{1i}'\beta_1 + u_{1i} \tag{10.1}$$

$$Y_{2i}^* = X_{2i}'\beta_2 + u_{2i} \tag{10.2}$$

其中 Y_{1i}^* 和 Y_{2i}^* 是(不可观测)潜变量。可观测的因变量为 Y_{1i}和 Y_{2i},定义为

$$Y_{1i} = \mathbf{1}(Y_{1i}^* > 0) \tag{10.3}$$

$$Y_{2i} = Y_{2i}^*\mathbf{1}(Y_{i1} = 1) \tag{10.4}$$

其中 $\mathbf{1}(A)$表示事件 A 的示性函数。我们只观测到 Y_{1i}^* 的

符号,只有在 $Y_{1i}=1(Y_{1i}^*>0)$时可以观测到 Y_{2i},而(X_{1i},X_{2i})是总可以观测到的。式(10.1)至式(10.4)在计量经济学文献中称为“类型 2”(Type-2)Tobit 模型。这种模型的一个典型应用出现在劳动力供给决策模型中,我们将以例子形式简单介绍。在劳动力供给模型中,若 $Y_{1i}=1$,这个个体进入劳动力队伍,Y_{1i}^*可以是最优劳动时间。若 $Y_{1i}^*\leqslant 0$,个体将继续保持失业。然而,即使个体参与劳动我们也观测不到 Y_{1i}^*的确切值。Y_{2i}为个体的(对数)工资,只有当个体参与工作$(Y_{1i}=1)$时才可观测到。协变量(X_{1i},X_{2i})包含参加工作的个体的可观测特征。

估计式(10.2)是很有趣的,记住只有当 $Y_{i1}=1$ 时,Y_{2i}才可观测。于是 Y_{2i}关于 $Y_{1i}=1$ 和 X_i 的条件期望为

$$E(Y_{2i}\mid X_i,Y_{1i}=1)=X_{2i}'\beta_2+E(u_{2i}\mid X_i,Y_{1i}=1) \tag{10.5}$$

联立式(10.2),有

$$Y_{2i}=X_{2i}'\beta_2+E(u_{2i}\mid X_i,Y_{1i}=1)+\varepsilon_{2i} \tag{10.6}$$

其中 $\varepsilon_{2i}=Y_{2i}-E(Y_{2i}\mid X_i,Y_{1i}=1)$,于是 $E(\varepsilon_{2i}\mid X_i,Y_{1i}=1)=0$。因此,若 $E(u_{2i}\mid X_i,Y_{1i}=1)$能够被一致地估计,则 β 可以从式(10.6)来估计。在参数框架下,通常假定(u_{1i},u_{2i})服从联合正态分布 $N(0,\Sigma)$,其中

$$\Sigma=\begin{bmatrix}\sigma_{11}, & \sigma_{12}\\ \sigma_{21}, & \sigma_{22}\end{bmatrix}$$

然后可以证明(见习题 10.1)

317

$$E(u_{2i}\mid X_i,Y_{1i}=1)=\sigma_{11}^{-1/2}\sigma_{12}\phi(X_{1i}'\beta_1/\sigma_{11}^{1/2})\Phi^{-1}(X_{1i}'\beta_1/\sigma_{11}^{1/2}) \tag{10.7}$$

其中 $\phi(\cdot)$和 $\Phi(\cdot)$分别是标准正态变量的概率密度函数和累积分布函数(Amemiya (1985,p.367))。将式(10.7)代入式(10.6)中得到的方程可以用来估计 β_2,其中可以使用 Heckman(1974,1979) 给出的二阶段程序,也可以使用最大似然方法。两种方法的详细比较见 Pagan 和 Ullah (1999,p. 303)。

10.2　半参数类型 2 Tobit 模型的估计

不对 u_{1i}与 u_{2i}的联合分布施加任何参数分布假定,假定 X_i 独立于(u_{1i},u_{2i}),可以得到如下结果:

$$E(u_{2i}\mid X_i,Y_{1i}=1)=E(u_{2i}\mid X_i,u_{1i}>-X_{1i}'\beta_1)=g(X_{1i}'\beta_1) \tag{10.8}$$

其中 $g(z)=E(u_{2i}\mid u_{1i}>-z)=1-F_{u_2\mid u_1}(-z)$,$F_{u_2\mid u_1}$是给定 u_{1i}时 u_{2i}的条件累积分布函数。因此 $g(\cdot)$的函数形式是未知的。将式(10.8)代入式(10.6)可得

$$Y_{2i}=X_{2i}'\beta_2+g(X_{1i}'\beta_1)+u_{2i} \tag{10.9}$$

式(10.9)是一个部分线性单一指标模型。Powell(1987) 提出了一个两步

估计过程来估计式(10.9)。首先用 Powell (1984)的 CLAD 估计量来估计 β_1，见 11.2 节，记为$\tilde{\beta}_1$；然后根据第 7 章中介绍的 Robinson (1998)的方法来估计 β_2，不过现在 $X_{1i}'\tilde{\beta}$是一个生成的回归元，因此，渐近方差与第 7 章中给出的结果不同。式(10.9)是 Ichimura 和 Lee(1991) 中半参数多指标模型的一个特例，Ai (1997)给出的一般半参数模型也包括了式(10.9)。然而，应用此方法时需要一定的识别条件。像第 7 章中讨论的部分线性模型那样，我们知道在式(10.9)中截距项是不可识别的，于是潜在的截距项要合并到 $g(\cdot)$中。另外，为了识别 β_1，这里还需要第 8 章中讨论的单一指标模型的识别条件。

令 $\theta=(\beta_1',\beta_2')'$，Ichimura 和 Lee(1991) 提出通过最小化以下目标函数来估计 θ

$$Q_n(\theta)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i\in\mathcal{X})[Y_i-X_{2i}'\beta_2-\hat{g}(X_i'\beta_1)]^2$$

其中

$$\hat{g}(X_{1i}'\beta_1)=\frac{\sum_{j=1}^{n}[Y_{2j}-X_{2j}'\beta_2]K_h[(X_{1i}-X_{1j})'\beta_1/h]}{\sum_{j=1}^{n}K_h[(X_{1i}-X_{1j})'\beta_1/h]}$$

这是 $E[Y_{2i}-X_{2i}'\beta_2\mid X_{1i}'\beta_1]$的一个核估计量，其中 $K_h(v)=h^{-1}K(v/h)$，$\mathcal{X}$是一个第 8 章中讨论过的修剪集。令$\hat{\theta}$为得到的 θ 的估计量，在类似于第 8 章给出的正则条件下，Ichimura 和 Lee 推导出了如下结果：

$$\sqrt{n}(\hat{\theta}-\theta)\xrightarrow{d}N(0,A^{-1}\Sigma A^{-1}) \tag{10.10}$$

其中 $A=E[\mathbf{1}(X_i\in\mathcal{X})\Delta(X_i)\Delta(X_i)']$，$\Sigma=E(u_{2i}^2B_iB_i')$，$B_i=\mathbf{1}(X_i\in\mathcal{X})-E[\mathbf{1}(X_i\in\mathcal{X})\mid X_{1i}'\beta_1]$，$\Delta(X_i)=X_{2i}[X_{2i}-E(X_{2i}\mid X_{1i}'\beta_1)]'+X_{1i}[X_{1i}-E(X_{1i}\mid X_{1i}'\beta_1)]'g(X_{1i}'\beta_1)$。Ichimura 和 Lee (1991)也给出了 A 和 Σ 的一致估计量。

10.2.1 Gallant 和 Nychka 的估计量

Gallant 和 Nychka (1987) 建议用级数展开来逼近联合密度$f(u_1,u_2)$，其中(u_1,u_2)是式(10.1)和式(10.2)所定义的误差项。他们提出用下式来逼近$f(\cdot,\cdot)$：

$$\tilde{f}(u_{1i},u_{2i})=\exp[-u_{1i}^2/(2\sigma_1^2)-u_{2i}^2/(2\sigma_2^2)]\left[\sum_{j=0}^{K}\sum_{k=0}^{K}\gamma_{jk}u_{1i}^ju_{2i}^k\right] \tag{10.11}$$

其思路是使用一个联合正态表达式作为基准分布，伴随着一个考虑到累积分布函数的一般形式的幂级数展开。应用以上联合密度公式可以计算出$f(u_{1i}\mid u_{2i})$且构造出一个对数似然函数，并可通过最大化对数似然函数来得到 β_1 和其他

参数的估计。Gallant 和 Nychka 在 $K\to\infty$ 和 $n\to\infty$ 时有 $K/n\to 0$ 的条件下证明了他们所提出的估计量的一致性。Coppejans 和 Gallant(2002) 证明了在估计 $f(u_1,u_2)$ 时可以使用数据驱动方法来选择幂级数展开项。　319

Newey(1999) 提出了一个替代的两阶段级数估计方法，可以先对 β_1 进行有效估计，再通过解一个有效得分方程来估计 β_2。Pagan 和 Ullah(1999, pp. 311—314) 提供了一个详细的关于 Newey 的级数估计方法及其渐近分布的讨论。我们将在第 15 章讨论一般级数方法。

10.2.2　选择模型中截距项的估计

如前面所讨论的，基于式(10.9)的估计方法由于不能将截距从 $g(\cdot)$ 中分离出来，故不能识别截距项。如果只关注 X_i 对 Y_{2i} 的局部效应，那么就可以不用考虑截距项。然而，在其他情形下，截距项可能也是非常重要的：决定参加工会的工人与未参加工会的工人之间的"工资缺口"时，分解不同社会经济团体的工资差异时，等等。令 μ 表示截距项，我们记 $X_{2i}=(1,\tilde{X}_{2i}')'$，$\beta_2=(\mu,\delta')'$。Heckman(1990) 建议对那些 $E(u_{2i}\mid Y_{1i}=1)=g(X_{1i}'\beta_1)$ 为零的观测值进行平均来估计截距项 μ。这样可以得到

$$\tilde{\mu}=\frac{\sum_{i=1}^{n}(Y_{2i}-\tilde{X}_{2i}'\hat{\delta})Y_{1i}\mathbf{1}(X_{1i}'\hat{\beta}_1>\gamma_n)}{\sum_{i=1}^{n}Y_{1i}\mathbf{1}(X_{1i}'\hat{\beta}_1>\gamma_n)}$$

其中 γ_n 是满足当 $n\to\infty$ 时 $\gamma_n\to\infty$ 的窗宽。

由于示性函数 $\mathbf{1}(\cdot)$ 是不可微的，如上定义的 $\tilde{\mu}$ 的渐近分布就很难验证。Andrews 和 Schafgans(1998) 建议用一平滑非减的累积分布函数来代替 $\tilde{\mu}$ 中的示性函数，其中该累积分布函数满足对 $z\leqslant 0$ 有 $s(z)=0$，对 $z\geqslant b$，$0<b<\infty$ 有 $s(z)=1$，且 $s(\cdot)$ 具有三阶有界导数。Andrews 和 Schafgans 提出用下式来估计 μ：

$$\hat{\mu}=\frac{\sum_{i=1}^{n}(Y_{2i}-\tilde{X}_{2i}'\hat{\delta})Y_{1i}S(X_{1i}'\hat{\beta}_1>\gamma_n)}{\sum_{i=1}^{n}Y_{1i}S(X_{1i}'\hat{\beta}_1>\gamma_n)}$$

Andrews 和 Schafgans(1998) 证明了一致性且推导出了 $\hat{\mu}$ 的渐近分布。$\hat{\mu}-\mu$ 的收敛速率(向零)依赖于 $v_i=X_{1i}'\beta_1$ 的分布。在某些情形下，收敛速率可以非常接近 $n^{-1/2}$。　320

其他用来估计半参数类型 2 Tobit 模型的方法包括 Newey (1991a,1999)基

于级数的估计量,Cosslett (1983)的逼近似然估计量,以及 Powell(1987)基于核的部分线性单一指标模型估计量。

10.3 半参数类型 3 Tobit 模型

本节我们研究下面 10.3.1 节中定义的类型 3 Tobit 模型的半参数估计。Vella(1992) 提出了一个使用类型 3 Tobit 残差作为一个生成回归元的样本选择偏差检验。使用类似的思路,Wooldridge(1994) 提出了一个比 Heckman(1979) 的方法更易于使用且更稳健的两阶段估计量。Li 和 Wooldrige (2002) 考虑到误差分布的非正态性,对 Wooldridge 的方法进行了一般化,得到了一个半参数部分线性模型,其中生成的回归元通过非参数形式进入模型。Chen (1997), Honoré, Kyriazidou 和 Udry (1997),以及 Lee (1994) 都提出了不需要误差分布的不同的半参数二阶段方法。

10.3.1 计量经济学预备

考虑通过潜变量定义的类型 3 Tobit 模型

$$Y_1^* = X_1'\beta_1 + u_1 \tag{10.12}$$

$$Y_2^* = X_2'\beta_2 + u_2 \tag{10.13}$$

其中第一个方程是选择方程,第二个方程是主要关注方程。只有当选择变量 Y_1^* 为正时因变量 Y_2^* 才可被观测到。因此我们观察 Y_1 与 Y_2 满足

$$Y_1 = \max\{Y_1^*, 0\} \tag{10.14}$$

$$Y_2 = Y_2^* \mathbf{1}(Y_1 > 0) \tag{10.15}$$

321 其中 Y_1 与 Y_2 是观测到的因变量,X_1 与 X_2 分别是具有维数 p_1 和 p_2 的外生变量向量,而 β_1 与 β_2 为相应的未知参数列向量。例如,在劳动力市场应用中,Y_1 会是一个个体的劳动时间,而 Y_2 会是(对数)小时工资率。同 10.1 节中讨论的类型 2 Tobit 模型相比,这里当劳动时间为正时可以观察到它,而在类型 2 Tobit 模型中只知道劳动时间为正或者为零,即不能观测到具体准确的劳动时间。

在式(10.14)与式(10.15)描述的选择规则下,我们有

$$E(Y_2^* \mid X_1, X_2, Y_1^* > 0) = X_2'\beta_2 + E(u_2 \mid u_1 > -X_1'\beta_1, X_1, X_2) \tag{10.16}$$

于是,若式(10.16)右端第二项非零时,用最小二乘方法关于 X_2 来回归 Y_2 可得到 β_2 的一个非一致的估计量。在 (u_1, u_2) 服从联合正态分布的假定下,Heckman(1976,1979)提出了一个简单的二阶段方法来估计类型 2 或者类型 3 Tobit 模型。Heckman 的提议为,通过加入一个选择偏差项的估计,在式(10.15)中恢复零条件均值,$E(u_2 \mid u_1 > -X_1'\beta_1, X_1, X_2)$。在正态情形下,此项

与逆 Mills 比成比例且只依赖于式(10.12)的未知参数,未知参数可以通过 Probit 或 Tobit 最大似然方法来估计。

Vella (1992,1993)和 Wooldrige (1994)提出不同的二阶段估计模型可能具有更好的有限样本性质。在(X_1,X_2)独立于(u_1,u_2)的假定下,他们注意到$E(u_2|x,u_1,Y_1>0)=E(u_2|u_1,Y_1>0)$。如果进一步假定$E(u_2|u_1)=\gamma_1 u_1$,那么选择偏差修正项为$\gamma_1 u_1$。可以用$\hat{u}_1=Y_1-X_1'\hat{\beta}_1$来估计$u_1$,其中$\hat{\beta}_1$为$\beta_1$的 Tobit 估计量。于是,可以使用$u_1$而非 Heckman (1979) 的逆 Mills 比,来作为条件期望中附加的变量。这样做的优点是即便X_2与逆 Mills 比接近共线时,u_1仍比X_2具有更多变化,因此使得 Vella-Wooldridge 估计量更稳定且更有效;更多细节见 Wooldridge(2002,p.573)。

并没有必要假定(u_1,u_2)的联合分布是已知的,或者假定$E(u_2|u_1)=\gamma_1 u_1$。当(u_1,u_2)的联合分布未知时,有$E(u_2|u_1)=g(u_1)$,其中$g(\cdot)$是一未知函数。此时容易证明$E(Y_{2i}|X_i,u_{1i})=X_{2i}'\beta_2+g(u_{1i})$。于是我们有

$$Y_{2i}=X_{2i}'\beta_2+g(u_{1i})+v_{2i} \qquad (10.17)$$

其中v_{2i}满足$E(v_{2i}|u_{1i},Y_{1i}>0)=0$。

322

依据 Robinson(1988) 并使用那些$Y_{1i}>0$的观测值,从式(10.17)我们得到

$$Y_{2i}-E(Y_{2i}|u_{1i})=[X_{2i}-E(X_2|u_{1i})]'\beta_2+v_{2i} \qquad (10.18)$$

或者是一个密度加权的相应结果

$$[Y_{2i}-E(Y_{2i}|u_{1i})]f_i=f_i[X_{2i}-E(X_2|u_{1i})]'\beta_2+f_i v_{2i} \qquad (10.19)$$

其中$f_i=f(u_{1i})$是u_1的概率密度函数。

Li 和 Wooldridge(2002) 给出了一个估计β_2的多步骤方法,过程如下:

(i) 用$\hat{u}_{1i}=Y_{1i}-X_{1i}\hat{\beta}_1$来估计$u_{1i}$,其中$\hat{\beta}_1$是$\beta_1$的第一阶段估计量。

(ii) 使用$\{Y_{2i},X_{2i},\hat{u}_{1i}\}_{i=1}^{n_1}$来得到$E(Y_{2i}|u_{1i})$和$E(X_{2i}|u_{1i})$的非参数核估计,记$\hat{Y}_{2i}=\hat{E}(Y_{2i}|\hat{u}_{1i})$和$\hat{X}_{2i}=\hat{E}(X_{2i}|\hat{u}_{1i})$。

(iii) 基于式(10.18) 应用最小二乘方法来估计β_2(如 Robinson (1988)),即,

$$\hat{\beta}_2=\left[\sum_{i=1}^{n}(X_{2i}-\hat{X}_{2i})(X_{2i}-\hat{X}_{2i})'\right]^{-1}\sum_{i=1}^{n}(X_{2i}-\hat{X}_{2i})(Y_{2i}-\hat{Y}_{2i})$$

我们假定在第一步,存在一个β_1的$\sqrt{n}$一致且渐近正态分布估计量,记为$\hat{\beta}_1$,由如下等式来刻画:

$$\sqrt{n}(\hat{\beta}_1-\beta_1)=\frac{1}{\sqrt{n}}\sum_{i=1}^{n}r_i+o_p(1)\rightarrow N(0,\Omega_2)\text{ 依分布收敛} \qquad (10.20)$$

其中$\Omega_2=E[r_i r_i']$,$\hat{r}_i$是r_i的一个估计量;于是$\hat{\Omega}_2=n^{-1}\sum_i \hat{r}_i\hat{r}_i'$是$\Omega_2$的一个一致估计量。

例如,Powell (1984) 的 CLAD 估计量满足式(10.20)并由下式定义:

$$\hat{\beta}_1 = \arg\min_{\beta_1} \frac{1}{n} \sum_{i=1}^{n} |Y_{1i} - \max\{0, X_{1i}'\beta_1\}| \tag{10.21}$$

323 对这种情形,r_i(在式(10.20)中定义)和$\hat{\Omega}$的特定定义在 Powell(1984)中给出。

Li 和 Wooldridge (2002) 证明了他们给出的 β_2 的估计量的$\sqrt{n}$正态性如下:

$$\sqrt{n}(\hat{\beta}_2 - \beta_2) \to N(0, \Sigma)\text{依分布收敛}$$

其中 $\Sigma = \Phi^{-1}[\Omega_1 + \Psi\Omega_2\Psi']\Phi^{-1}$, $\Phi = E[v_i'v_if_i^2]$, $v_i = X_{2i} - \xi_i$, $\xi_i = E(X_{2i}|u_{1i})$, $\Psi = E\{v_if^2(u_{1i})g^{(1)}(u_{1i})[X_{1i} - E(X_{1i}|u_{1i})]\}$, $\Omega_1 = E[v_i'v_i\sigma_u^2(X_{2i}, u_{1i})f_i^4]$, $u_{1i} = Y_{1i} - X_{1i}'\beta_1$, $g^{(1)}(u_1) = \mathrm{d}g(u_1)/\mathrm{d}u_1$, Ω_2 在式(10.20)中给出。

Li 和 Wooldridge (2002) 也给出了渐近方差 Σ 的一个一致估计量,并进一步导出了此模型的半参数有效性边界。上述多步骤方法在 Σ 与半参数下界相差一个正定矩阵的意义下不是半参数有效的。有效估计通常可以通过一个一步过程来得到,如 Ai(1997)中 β_1 与 β_2 是同步被估计的。

在上面列出的方法中,生成的回归元是从参数模型得到的。Ahn 和 Powell (1993) 考虑了一个更一般的情形,其中生成的回归元是从非参数回归模型得到的。他们的方法也可以用来在半参数类型 3 Tobit 模型估计 β_2。

10.3.2 不同的估计方法

许多作者也提出了在不需要知道(u_1, u_2)的联合分布的相关信息下类型 3 Tobit 模型的半参数估计;见 Chen (1997),Honoré 等. (1997),Lee(1994),以及其他。下面,我们简要讨论几个这种估计量。

Chen(1997)观察到在(u_1, u_2)独立于(X_1, X_2)的条件下,

$$E(Y_2|X_1, X_2, u_1 > 0, X_1'\beta_1 > 0, Y_1 > 0) =$$
$$E(Y_2|u_1 > 0, x) = X_2'\beta_2 + \alpha_0 \tag{10.22}$$

其中 α_0 是一个常数。然而,α_0 并不是原始模型的截距,这是由于没有进一步的

324 假定,截距项是不可识别的。基于式(10.22),Chen 建议用一个应用于修剪过的子样本的简单最小二乘过程来估计 β_2,

$$\hat{\beta}_{2,\text{Chen}} = \arg\min_{\beta_2, \alpha} \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{Y_{1i} - X_{1i}'\hat{\beta}_1 > 0, X_{1i}'\hat{\beta}_1 > 0\}} (Y_{2i} - X_{2i}'\beta_2 - \alpha)^2 \tag{10.23}$$

其中$\hat{\beta}_1$ 是第一步中 β_1 的$\sqrt{n}$一致估计量,即由 Honoré 和 Powell(1994)给出的估计量,或 Powell(1984)的 CLAD 估计量。如 Chen 所讨论的,式(10.23)给出的估计量存在的一个问题是,它可能剔除过多的观察测值而导出一个无效估计。Chen 因此提出一个替代的估计量,其在有限样本应用中只剔除很少的数据点

(详见 Chen(1997)中的式(11))。

Honoré 等 (1997) 考虑了一个不同的方法。为了放松 Heckman (1979) 的正态假定,Honoré 等考虑了潜在误差项条件于回归元是对称分布的情形,其中允许存在任意的异方差。在此情形下,样本选择的影响为,在样本选择的条件下,误差项不再是对称分布的。Honoré 等注意到如果用满足 $-X_1'\beta_1 < u_1 < X_1'\beta_1$ (等价于 $0 < Y_1 < 2X_1'\beta_1$)的观测值来估计 β_2,那么 u_2 就是关于零对称分布的。于是下面最小绝对离差估计量可以一致地估计 β_2:

$$\hat{\beta}_{2,\mathrm{HKU}} = \arg\min_{\beta_2} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{0 < Y_{1i} < 2X_{1i}'\hat{\beta}_1\}} \left| Y_{2i} - X_{2i}'\beta_2 \right| \tag{10.24}$$

其中$\hat{\beta}_1$ 是 β_1 第一阶段的$\sqrt{n}$一致估计量,如 Powell(1984) 的 CLAD 估计量。Honoré 等也证明了他们提出的估计量$\hat{\beta}_{2,\mathrm{HKU}}$的$\sqrt{n}$正态性。

在误差项和回归元相互独立的假定下,Lee(1994,式(2.12))证明了

$$\begin{aligned} Y_{2i} - E(Y_2 \mid u_1 > -X_{1i}'\beta_1, X_1'\beta > X_{1i}'\beta_1) \\ = [X_{2i}' - E(X_2' \mid X_1'\beta_1 > X_{1i}'\beta_1)]\beta_2 + u_{2i} \end{aligned} \tag{10.25}$$

其中 u_{2i}满足 $E(u_{2i} \mid u_1 > -X_{1i}'\beta_1, X_i' \ \beta > X_{1i}'\beta_1) = 0$。Lee 建议首先用核估计量替换式(10.25)中的条件期望(同样,β_1 需要用一个第一步估计量来替换),再应用最小二乘过程来估计 β_2(我们定义为$\hat{\beta}_{2,\mathrm{Lee}}$)。Lee 证明了$\hat{\beta}_{2,\mathrm{Lee}}$的渐近正态性。

Chen(1997)和 Honoré 等(1997) 的方法均不需要非参数技术,而 Li 和 Wooldridge(2002)以及 Lee(1994)均使用了非参数核方法。尽管非参数核方法一般都是对平滑参数的选择敏感的,但 Lee, Min, Sheu 和 Wang(2003)中的 Monte Carlo 模拟意味着 Lee 与 Li 和 Wooldridge 的估计量对于平滑参数的选择是相当不敏感的。这是因为 β_2 的半参数估计量依赖于非参数估计量的平均值,而一个平均的非参数估计量与一个逐点的非参数估计量相比,对于不同值的平滑参数是更不敏感的。 325

在类型 2 Tobit 模型中也可以应用 Ichimura(1993) 基于单一指标的半参数非线性最小二乘方法来估计 β_1,如二元劳动力参与变量。只使用满足 $Y_{1i} > 0$ 的数据,相应的半参数工资方程是一个部分线性单一指标模型(见 Ichimura 和 Lee (1991)),具有以下形式:

$$Y_{2i} = X_{2i}'\beta_2 + \theta(X_{1i}'\beta_1) + \eta_{2i} \tag{10.26}$$

其中 $\theta(X_{1i}'\beta_1) = E(u_2 \mid u_1 > -X_{1i}'\beta_1)$具有未知的函数形式,$\eta_{2i}$ 满足条件 $E(\eta_{2i} \mid X_i) = 0$。Ichimura 和 Lee 提出了一个半参数非线性最小二乘方法来估计

式(10.26),且他们证明了他们所提出的统计量的渐近分布。

我们推荐有兴趣的读者参考 Christofides, Li, Liu 和 Min (2003),其中考虑了在应用方面五个半参数估计量的相对表现。这些估计量是(i) Chen(1997)的半参数估计量,(ii) Honoré 等(1997)的半参数估计量,(iii) Lee(1994)的半参数估计量,(iv) Li 和 Wooldridge(2002)的半参数估计量,(v)基于 Ichimura (1993)以及 Ichimura 和 Lee(1991)的半参数类型 2 Tobit 估计量。注意 Honoré 等要求 u_2 具有(条件)对称分布,但是不要求 (u_1,u_2) 独立于 (X_1,X_2);相反,其余四个估计量假定 (u_1,u_2) 独立于 (X_1,X_2) 而不要求 u_2 是对称分布。对称条件既不弱于也不强于独立条件。

326

对于选择偏误的检验,我们考虑检验无选择偏误,或者如 Vellla(1992)和 Wooldridge(1994)所描述的对参数选择偏误进行检验,这与在 Li 和 Wooldridge (2002)中所介绍的一般的半参数选择偏误相反。令 H_0^a 表示没有选择偏误的原假定。若 H_0^a 被拒绝,检验参数选择模型是否适合就非常必要,即是否有 H_0^b: $E(u_2|u_1)=u_1\gamma$ 几乎处处满足。如果误差是正态分布的,那么有 $g(u_1)=u_1\gamma$,且可以通过检验是否有 $\gamma=0$ 来对无选择偏误假定进行检验。然而,当 $g(u_1)\neq u_1\gamma$ 时,基于检验 $\gamma=0$ 的对无选择偏误的参数检验会给出误导的结果,这是由于此时会出现两类错误:第一类,H_0^a 为真时,当 $g(u_1)\neq u_1\gamma$ 时参数假定可能会拒绝原假定;第二类,当 H_0^a 非真时,参数假定就没有功效,即使样本容量趋于无穷大,这是因为这不再是一个一致检验(见第 12 章关于一致检验的定义)。

我们下述的检验统计量对 (u_1,u_2) 的分布假定是稳健的。也就是说,无论 (u_1,u_2) 的联合分布是什么,如果存在选择偏误,检验出选择偏误的概率当样本容量趋近于无穷时将收敛于 1。不存在选择偏误 (H_0^a) 的原假定可以表示为 $E(u_2|u_1)=0$。对立假定 (H_1^a) 可以表示为 $E(u_2|u_1)\equiv g(u_1)\neq 0$。若 H_0^a 为真,那么在 H_0^a 下,对可观测到的 Y_2 关于 X_2 的普通最小二乘回归给出了 β_2 的一个一致估计量(我们记为 $\hat{\beta}_{2,\text{ols}}$),同时,最小二乘残差 $\hat{u}_{2i}=Y_{2i}-X'_{2i}\hat{\beta}_{2,\text{ols}}$ 是 u_{2i} 的一个一致估计量(在 H_0^a 下)。像 Li 和 Wang(1998)以及 Zheng(1996)提出的参数模型设定的检验统计量那样(有关此类检验更详细的讨论见第 12 章),对 H_0^a 的一个检验统计量如下:

$$\hat{I}_n^a=\frac{1}{n_1(n_1-1)}\sum_{i=1}^{n_1}\sum_{j\neq i,j=1}^{n_1}\hat{u}_{2i}\,\hat{u}_{2j}K_h(\hat{u}_{1i}-\hat{u}_{1j}) \tag{10.27}$$

其中 n_1 表示可观测样本(即满足 $Y_{2i}>0$ 的观测值)的样本容量且 $\hat{u}_{1i}=Y_{1i}-X'_{1i}\hat{\beta}_1$。

我们现在给出一些正则条件,在这些条件下可以导出 $\tilde{I}_n^a$ 以及下面定义的另一检验 $\tilde{I}_n^b$ 的渐近分布。

条件 10.1 $(Y_{2i},X_i,u_{1i},u_{2i})$是独立同分布的,分布同于$(Y_2,x,u_1,u_2)$。$x$,$u_1$,$u_2$ 均具有有限的四阶矩。$\partial g(u_1)/\partial u_1$,$\partial^2 g(u_1)/\partial u_1^2$ 对 u_1 都是连续的且都被一个具有有限二阶矩的函数(记为 $M(u_1)$)占优。$\hat{\beta}_1-\beta_1=O_p(n^{-1/2})$。

条件 10.2 核函数 $K(\cdot)$是有界、对称、三次可微且具有有界导函数的。$\int K(v)\mathrm{d}v=1$,$\int K(v)v^4\mathrm{d}v<\infty$。 327

条件 10.3 当 $n_1\to\infty$ 时,$h\to 0$,$n_1h\to\infty$。

借鉴第12章中定理12.1的证明,我们可以证明

命题 10.1 在条件(10.1)至条件(10.3)之下,我们有(当 $n_1\to\infty$ 时)

(i) 若 H_0^a 为真,

$$n_1h^{1/2}\tilde{I}_n^a/\hat{\sigma}_a\xrightarrow{d}N(0,1)$$

(ii) 若 H_1^a 为真,

$$P[|n_1h^{1/2}\tilde{I}_n^a/\hat{\sigma}_a|>C]\to 1,\text{任意 } C>0$$

其中 $\hat{\sigma}_a^2=\dfrac{2h}{n_1(n_1-1)}\sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1}\hat{u}_{2i}^2\hat{u}_{2j}^2K_h^2(\hat{u}_{1i}-\hat{u}_{1j})$。

如果 H_0^a 被拒绝,就需要估计一个参数模型或半参数选择模型。于是,检验参数模型是否适合就十分重要。参数设定正确的原假定可以表示为 H_0^b:$E(Y_2|X_2,u_1)=X_2'\beta_2+u_1\gamma$。则对立假定为 $E(Y_2|X_2,u_1)=X_2'\beta_2+g(u_1)$,其中 $g(u_1)\neq u_1\gamma$。因此,检验是一个线性回归模型还是一个部分线性回归模型是非常必要的。Li 和 Wang(1998) 提出了当 u_1 可观测时的这样一个检验。在 Li 和 Wang 给出的检验中,将 u_{1i}替换为$\hat{u}_{1i}=Y_{1i}-X_{1i}'\hat{\beta}_1$,可得到一个检验 H_0^b 与 H_1^b 的有效方法。令 $\tilde{u}_{2i}=Y_{2i}-X_{2i}'\hat{\beta}_2-\hat{u}_{1i}\hat{\gamma}$,其中$\hat{\beta}_2$ 是 Li 和 Wooldridge(2002) 给出的 β_2 的半参数估计量,$\hat{\gamma}$是基于 $Y_{2i}=X_{2i}'\beta_2+\hat{u}_{1i}\gamma+\text{error}$ 得到的 γ 的普通最小二乘估计量。这时,检验统计量给出为

$$\tilde{I}_n^b=\frac{1}{n_1(n_1-1)}\sum_{i=1}^{n_1}\sum_{j\neq i,j=1}^{n_1}\tilde{u}_{2i}\tilde{u}_{2j}K_h(\hat{u}_{1i}-\hat{u}_{1j})$$

命题 10.2 在条件(10.1)至条件(10.3)下,我们有(当 $n_1\to\infty$ 时)

(i) 若 H_0^b 为真,$n_1h^{1/2}\tilde{I}_n^b/\hat{\sigma}_b\xrightarrow{d}N(0,1)$。

(ii) 若 H_1^b 为真,对任意的 $c>0$,$P[|n_1h^{1/2}\tilde{I}_n^b/\hat{\sigma}_b|>c]\to 1$,其中 $\hat{\sigma}_b^2=\dfrac{2h}{n_1(n_1-1)}\sum_{i=1}^{n_1}\sum_{j\neq i}^{n_1}\tilde{u}_{2i}^2\tilde{u}_{2j}^2K_h^2(\hat{u}_{1i}-\hat{u}_{1j})$。 328

命题10.2的证明同第12章中定理12.1的证明类似,这里就不作介绍。

注意$\tilde{I}_n^a$和$\tilde{I}_n^b$均只涉及一维的核估计，因此都避免了维数诅咒。当处理大的数据集时，检验统计量$\tilde{I}_n^a$和$\tilde{I}_n^b$提供了检测可能的样本选择偏误及检测是否需要用半参数选择模型来修正这个偏误的非常有力的途径。

应当提到的是，上述检验是被设计用来在模型是线性且可加的良好假定下，检验不存在选择偏误或参数选择偏误的。如果模型为线性且可加的假定不再成立，$\tilde{I}_n^a$和$\tilde{I}_n^b$检验可能由于这些假定被违背而拒绝原假定。理想地，应该进一步检验是半参数选择模型还是一般的非参数模型，它不依赖线性及可加性。然而，这样一个检验很可能遭受维数诅咒。

10.4 Das, Newey 和 Vella 的非参数选择模型

Das, Newey 和 Vella(2003) 考虑了下述非参数样本选择问题。令 Y 为因变量，X 为右侧的关注变量向量。假定有

$$\begin{aligned} Y_i^* &= g(X_i) + \varepsilon_i \\ Y_i &= d_i Y_i^* \end{aligned} \qquad (10.28)$$

其中 $g(\cdot)$的函数形式未指定，ε_i 为一扰动项，d_i 为二元选择指标。通过允许 $g(\cdot)$具有未知函数形式而非 X_i 的线性函数可以将常用的样本选择模型一般化。若 d_i 与 ε_i 是相关的，为了估计 $g(\cdot)$，此时需要对选择偏误进行修正。令 Z_i 为一决定选择的变量向量，定义 $p_i \overset{\text{def}}{=} E(d_i|X_i, Z_i)$（倾向得分）。Das 等做出了如下假定。

329 **假定 10.1**

(i) $E(\varepsilon|X, Z, d=1) = m(p)$。

(ii) 对任意随机变量 $l(X)$ 与 $b(p)$，$P[l(X) + b(p) = 0 | d=1] = 1$ 意味着 $l(x)$ 是一个常数。

Das 等(2003)证明了若 $d = \mathbf{1}(\alpha_0(Z) - \eta > 0)$（无需指定 $\alpha_0(\cdot)$的函数形式）且(ε, η)独立于(Z, X)，那么假定 10.1 成立。假定 10.1 中(i)立即意味着

$$E(Y|X, Z, d=1) = g(X) + m(p) \qquad (10.29)$$

式(10.29)设定 $g(\cdot)$为一可加的常数。在可加回归模型中，通过在一个可加回归模型中用选择（可观测的）的 Y 对 X 和 p 进行非参数回归可以估计$g(\cdot)$。因此，Das 等(2003)提出了一个基于级数的两步估计方法来估计$g(\cdot)$。第一步通过对倾向得分 $p = E(d|Z, X)$进行非参数估计来估计 p（记为$\hat{P}$），第二步则在施加了可加约束后关于 X 和$\hat{P}$对可观测的 Y 进行回归。之所以使用非参数级数方法

是因为很容易施加可加性约束（对级数方法的详细讨论见第15章）。同样，也可以用核方法来估计可加模型(10.29)，如边际积分法或者反向拟合法。

考虑到内生的 X，通过使用 Newey，Powell 和 Vella(1999)给出的方法，Das等(2003)进一步扩展了他们的选择模型(10.28)。这里，我们划分 $X=(Z_1,X_2)$ 和 $Z=(Z_1,Z_2)$。模型给出为

$$\begin{aligned} Y_i^* &= g(Z_{1i},X_{2i})+\varepsilon_i \\ X_{2i} &= \pi(Z_{1i},Z_{2i})+v_i \\ Y_i &= \mathrm{d}_i Y_i^* \end{aligned} \tag{10.30}$$

此时，假定10.1被修改如下：

假定10.2

(i) $E(\varepsilon|v,Z,d=1)=m(p,v)$。

(ii) 对任意的随机变量 $l(X)$ 和 $b(p,v)$，$P[l(X)+b(p,v)=0|d=1]=1$ 意味着 $l(x)$ 是一个常数。

在假定10.2下，Das等(2003)证明了 330

$$E(Y|X,Z,d=1)=g(X)+m(p,v) \tag{10.31}$$

式(10.31)给出了估计 $g(\cdot)$ 的基础。首先，基于 $E(d|X,Z)$ 和 $E(X_2|Z)$ 的非参数估计量来分别估计 p 和 v（由于 $p=E(d|X,Z)$，$v=X_2-E(X|Z)$）。在施加了可加约束后关于 X 和 (p,v) 对可观测的 Y 进行非参数回归。然后，由此估计 $g(\cdot)$（与 m 一起），即使用 Y、X 以及 v 与 p 的估计值来估计式(10.31)。Das等(2003)证明了所得到的估计量的渐近分布。然后他们应用他们提出的方法估计了受教育程度对澳大利亚年轻人工资的影响。

10.5 习 题

习题10.1 推导式(10.7)。

习题10.2 推导式(10.17)。

习题10.3 考虑式(10.27)中定义的 $\tilde{I}_n^a$。

(i) 证明在 H_0^a 下，$\tilde{I}_n^a \xrightarrow{p} 0$。

(ii) 证明 $\alpha\partial C>0$，$\tilde{I}_n^a \xrightarrow{p} C$ 是一个正的常数。

(iii) 直观讨论为什么 H_0^a 为伪时，$\tilde{I}_n^a$ 被期望具有功效。

提示：检验统计量的功效定义为 H_0 为伪时该检验拒绝 H_0 的概率。

第11章 截断模型

331

经济学家收集的数据往往在某些方面是不完备的。本章我们考虑这样一种情况,其数据被称为"截断的"(censored)。我们不能观察到一个因变量的值高于或低于某一特殊值时,截断才会发生,尽管我们确实观测到了相关的解释变量。也就是说,因变量的观测值在某一区域外被截断,在此区域的端点处存在多个观测值。在这类情形下,我们知道取零值因变量的观测数目和这些观测值对应的解释变量的取值。[①] 严格来讲,一个样本被截断是指没有观测值被系统地排除掉,但是一些信息被隐藏了。想象一个审查者阅读你的邮件并用墨涂掉其中的一部分——你仍然得到了你的邮件,尽管它的某些部分难以辨认。

在许多情形下都会产生截断数据。一个常见的截断的来源是"编码上限"(top coding)。例如,在可公开获得的 CPS 数据中就存在这种情况。对于可公开获得的数据,美国人口普查局当前使用的方法是在 99 999 美元每年处遮掉收入分布的顶部。于是,所有超过这个水平的收入者的收入在 CPS 公用数据库中显示为 99 999 美元每年,而不管他们实际收入为多少。另一个例子来自 Long(1997,p. 189),涉及一个模型,其因变量是科学家们第一个学术工作的声望,声望被划分在区间[1.0,5.0]中的一个连续标度上。然而,对于评级低于 1.0 声望水平的研究生项目或者没有研究生项目的系,工作的声望是不可得到的。无论这些实际的声望是多少,它们都以 1.0 的声望被记录。

332

我们首先考虑一个常用的传统参数方法来处理截断问题的出现,这为在过去 20 年间提出的许多具有创

① 断尾(truncation)作为另一种情况,发生在位于某区间外的因变量和解释变量的观测值都完全丢失时。

新性的半参数和非参数模型提供了背景。

11.1 参数截断模型

参数方法处理截断的出现是通过一个调整机制进行的，该机制能够克服直接应用最小二乘方法带来的偏误。这通常都伴随着对模型的残差规定一个特殊的概率密度函数，这样得到一个修正项，当它被一致估计时，能够得出所关注参数的一致估计。我们简要回顾一下这个流行的方法，然后为文献中出现的许多半参数和非参数方法做好铺垫。

考虑"潜变量模型"

$$Y_i^* = X_i'\beta + \varepsilon_i, \quad i = 1, \cdots, n \tag{11.1}$$

其中 $X_i \in \mathbb{R}^q$，β 是一个 $q\times1$ 参数向量，ε_i 为一均值为零的扰动项。然而，我们没有观测到 Y_i^*，而是观测到 Y_i 如下：

$$Y_i = \begin{cases} Y_i^*, & Y_i^* > 0 \\ 0, & Y_i^* \leqslant 0 \end{cases} \quad i = 1, \cdots, n \tag{11.2}$$

这被称为一简单的"左截断机制"[①]，即

$$Y_i = \max\{X_i'\beta + \varepsilon_i, 0\} \tag{11.3}$$

此模型易修改为容许右截断或左右同时截断机制，也可以通过从 Y_1 和 $X_i'\beta$ 中减去 c 变为容许非零截断点机制。对下述讨论，我们假定使用的是左截断机制。 333

我们在式(11.3)下进行讨论，即不失一般性，Y_i 在零处截断，于是 ε_i 在 $-X_i'\beta$处截断（即，$\varepsilon_i > -X_i'\beta$）。很清楚地，$Y_i$ 的均值不同于 Y_i^* 的均值。对于截断数据，Y_i 的变动将会低估真实值 Y_i^* 的变动，于是，在应用经典最小二乘方法时，一般来说，得到的参数估计是存在向零的偏误。

处理截断模型最简单的传统参数估计方法是通过似然法来估计 β，这需要为 u_i 指定一个具体的密度函数，最经常使用的是正态分布。一个流行的方法来自 Heckman(1979)，他提出了一个只使用涉及 $Y_i > 0$ 的观测值的两步过程。给定 n_1 个 $Y_i^* > 0$ 的观测值，其中 $n_1 < n$，关于这些观测值的回归方程为（当 $Y_i^* > 0$ 时，$Y_i = Y_i^*$）

$$Y_i = X_i'\beta + \varepsilon_i, \ i = 1, 2, \cdots, n_1$$

现在考虑给定 $Y_i^* > 0$ 时 Y_i 的条件期望，即

---

① 此模型有时被称为"Tobit 模型"（Tobin (1958)）。Tobin 注意到支出是非连续的，只有"渴望"超出了某一水平时才会进行购买，并建立了家庭耐用品支出的模型。因此，只有确实购买了商品支出才不为零，未进行购买的家庭则支出记为零。也可参考 Heckman (1976,1979)。

$$E[Y_i|X_i,Y_i>0]=X_i'\beta+E[\varepsilon_i|Y_i>0]$$
$$=X_i'\beta+E[\varepsilon_i|\varepsilon_i>-X_i'\beta]$$

假定 $\varepsilon_i \sim N(0,\sigma^2)$，扰动项 ε_i 条件于 X_i 和 $Y_i^*>0$（即 $\varepsilon_i>-X_i'\beta$）的条件概率密度函数给出如下

$$f(\varepsilon_i|\varepsilon_i>-X_i'\beta)=\frac{\phi(\varepsilon_i/\sigma)}{1-\Phi(-X_i'\beta/\sigma)},\varepsilon_i>-X_i'\beta$$

其中 ϕ 和 Φ 是标准正态随机变量的概率密度函数和累积分布函数。于是，ε_i 的条件期望为

$$E[\varepsilon_i|\varepsilon_i>-X_i'\beta]=\int_{-X_i'\beta/\sigma}^{\infty}\frac{\varepsilon f(\varepsilon)}{1-F(-X_i'\beta/\sigma)}\mathrm{d}\varepsilon$$
$$=\sigma\frac{\phi(X_i'\beta/\sigma)}{\Phi(X_i'\beta/\sigma)} \tag{11.4}$$
$$=\sigma\lambda_i$$

其中 $\lambda_i=\phi(X_i'\beta/\sigma)/\Phi(X_i'\beta/\sigma)$。

考虑到 ε_i 的均值非零，于是关于 m 个 $Y_i^*>0$ 的观测值的回归方程可写成

$$Y_i=X_i'\beta+\sigma\lambda_i+u_i,\ i=1,2,\cdots,n_1$$

334 其中，显然由于 $E(Y_i|X_i,Y_i>0)=X_i'\beta+\sigma\lambda_i$，有 $E(u_i|X_i,Y_i>0)=0$。若 λ_i 可观测，则可以应用最小二乘方法来估计 β（和 σ），且得到的估计量是无偏的。然而，即使 λ_i 不可观测，给定正确的参数设定的条件下，可以得到一致估计，这就引出 Heckman(1979) 著名的两步估计过程，其中 λ_i 通常被称为"Heckman 的 λ"。Heckman 的方法是使用一个 Probit 模型，有观测值被截断则结果为 0，否则为 1，通过极大似然方法首先估计出不可观测项 λ_i，记为$\hat{\lambda}_i$。接下来只使用满足 $Y_i^*>0$ 的观测值来估计 $Y_i=X_i'\beta+\sigma\hat{\lambda}_i+\varepsilon_i$（详见 Amemiya(1985)）。若我们令 $\gamma=(\beta,\sigma)'$，$\hat{Z}_i=(X_i,\hat{\lambda}_i)$，这就产生了一个具有以下形式的最小二乘估计量：

$$\hat{\gamma}=\begin{bmatrix}\hat{\beta}\\ \hat{\sigma}\end{bmatrix}=\left(\frac{1}{n_1}\sum_{i=1}^{n_1}\hat{Z}_i\hat{Z}_i'\right)^{-1}\frac{1}{n_1}\sum_{i=1}^{n_1}\hat{Z}_iY_i$$

显然$\hat{\beta}$的性质依赖于回归方程和 ε_i 密度的正确设定。当 ε_i 的密度被错误设定时，其条件均值本身将被误设从而导致 β 的非一致估计。

当 ε_i 具有异方差时，最大似然估计量的潜在不一致性被广泛研究（见 Arabmazar和 Schmidt(1981)以及其中的参考文献），而经典回归模型中正态条件下的最大似然估计量当然通常是一致的。

Powell(1984)提出了一个流行的半参数方法，其对存在有条件异方差以及

ε_i 分布的错误设定是稳健的,我们将在 11.2 节进行讨论。Powell 的估计量是一个 L_1 范数方法(norm approach),他命名为"截断最小绝对离差"(censored least absolute deviations,CLAD)。尽管比较流行,但是当截断很严重时,此估计量就会失效。一个Chen 和 Khan(2000)提出的基于核的解决方法在 11.3 节给出。

我们现在考虑大量最近发展起来的用来处理截断数据的半参数估计量。

11.2 半参数截断回归模型

Powell(1984)的截断最小绝对离差(CLAD)估计量使用了一个与 Heckman (1979) 显著不同的方法。Powell 考虑了最小绝对方差的 L_1 范数方法而非使用最小二乘的 L_2 范数方法,从这样的观测开始,要观察到 Y_i^* 的真实值,在误差项 ε_i 的中位数为零的假定下,其中位数应与回归模型 $X_i'\beta$ 重合。在数据为左截断的情形下,例如零为截断点,那么中位数依然是重合的,如果对样本中一个正的部分的个体,真实因变量 Y_i^* 的典型值(即中位数)非负,它仍然成立,对很多总体这都是合理的,但对于截断严重的总体可能会不再成立。也就是说,当 Y_i 被截断,若回归函数 $X_i'\beta$ 位于未截断区间中,则 Y_i 的中位数不受截断出现的影响,否则 Y_i 的中位数将是截断点。 335

利用 ε_i 条件中位数为零的约束,Powell (1984) 的 CLAD 估计量定义为使下式最小化的变量:

$$S_n(\beta)=\frac{1}{n}\sum_{i=1}^{n}|Y_i-\max\{X_i'\beta,0\}| \tag{11.5}$$

易知式(11.5)等价于最小化

$$S_n(\beta)=\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(X_i'\beta>0)|Y_i-X_i'\beta| \tag{11.6}$$

这是由于,对于 $X_i'\beta<0$ 的观测值,我们有 $\max\{X_i'\beta,0\}=0$,于是 $|Y_i-\max\{X_i'\beta,0\}|=|Y_i|$,而不依赖于 β。令$\hat{\beta}$表示最小化式(11.6)得到的 β 的估计量,Powell 证明了如下结果:

$$\sqrt{n}(\hat{\beta}-\beta)\xrightarrow{d}N(0,V_{\mathrm{CLAD}}^{-1})$$

其中 $V_{\mathrm{CLAD}}=4f^2(0)E[\mathbf{1}(X_i'\beta>0)X_iX_i']$,其中 $f(0)$ 为 ε_i 在原点处的密度。假定 ε_i 独立于 X_i,那么 $f(0)=\lim\limits_{h\to0}P(0\leqslant\varepsilon_i<h)=\lim\limits_{h\to0}P(0\leqslant\varepsilon_i<h\,|\,X_i'\beta>0)$。于是,Powell 建议使用下式来估计 $f(0)$:

$$\hat{f}(0)=\frac{\mathbf{1}(X_i'\hat{\beta}>0)\mathbf{1}(0\leqslant\hat{\varepsilon}_i<h)}{h\sum_{i=1}^{n}\mathbf{1}(X_i'\hat{\beta}>0)}$$

336 Horowitz 和 Neumann(1987)提出了一个不同的估计量,使用满足 $X_i'\hat{\beta}\in[-h/2,h/2]$ 的数据来估计 $f(0)$:

$$\hat{f}(0)=\frac{\sum_{i=1}^{n}\mathbf{1}\left(-\frac{h}{2}\leqslant\hat{\varepsilon}_i\leqslant\frac{h}{2}\right)\mathbf{1}(Y_i>0)}{h\left[\sum_{i=1}^{n}\mathbf{1}\left(X_1'\hat{\beta}>\frac{h}{2}\right)+\frac{1}{2}\left(1+\frac{X_1'\hat{\beta}}{(h/2)}\right)\mathbf{1}\left(-\frac{h}{2}<X_i'\hat{\beta}\leqslant\frac{h}{2}\right)\right]} \tag{11.7}$$

Hall 和 Horowitz (1990) 建议用一个核函数来代替示性函数。Newey 和 Powell (1990) 修正了目标函数,得到

$$\sum_{i=1}^{n}w_i\left|Y_i-\max\{X_i'\beta,0\}\right|$$

并且证明了最佳的权重应为 $w_i=2f(0|X_i)$。得到的估计量的渐近方差可以证明为 $\{4E[\mathbf{1}(X_i'\beta>0)f^2(0|X_i)X_iX_i']\}^{-1}$,并且 Newey 和 Powell 证明对在条件中位数为零的约束下的截断回归模型,该渐近方差等于半参数有效边界。当然,若 ε_i 独立于 X_i,则 $f(0|X_i)$ 为一常数 $f(0)$,于是 Newey 和 Powell 的估计量退化为$\hat{\beta}$。

Powell(1986)也提出了一个 β 的对称截断最小二乘估计量。假定 ε_i 关于零对称分布,那么 ε_i 具有非零的条件均值,这是由于满足 $\varepsilon_i>-X_i'\beta$ 的观测值是被截断的;若同时也去掉 $\varepsilon_i<X_i'\beta$ 时的观测值,那么 ε_i 的条件均值就又变为零。因此,可以用满足 $\varepsilon_i\in[-X_i'\beta,X_i'\beta]$ 或等价地满足 $Y_i\in[0,2X_i'\beta]$ 的观测值来一致地估计 β。其他的估计方法有 Newey (1991b)提出的基于广义矩的估计方法,利用误差项的条件对称性来得到有效估计量,再基于其有效得分来估计 β,以及 Honoré 和 Powell (1994) 给出的同截断最小绝对离差(identically censored least absolute deviation,ICLAD)估计量和同截断最小二乘(identically censored least squares,ICLS)估计量。

11.3 具有非参数异方差的半参数截断回归模型

Chen 和 Khan(2000)考虑了关于异方差截断线性回归模型的估计过程,它
337 对识别条件的要求比 Powell(1984)的 CLAD 估计量要弱,且允许各种程度的截断。Chen 和 Khan 观察到当矩阵 $E[\mathbf{1}(X_i'\beta>0)X_iX_i']$ 不是满秩时使用 CLAD 估计量会产生严重的经验问题,于是 β 不可识别。这在指标 $X_i'\beta$ 以大概率取负值的典型情形下产生,数据截断严重时也经常出现。Chen 和 Khan 通过将条件异方差限制为乘积形式,允许比 Powell 的 CLAD 估计量更弱严格的识别条件。他

们将误差项设定为一个同方差的误差项与可用核方法估计的回归元标量函数的乘积。他们假定

$$\varepsilon_i = \sigma(X_i)v_i, P(v_i \leqslant \lambda \mid X_i) \equiv P(v_i \leqslant \lambda)\ \forall \lambda \in \mathbb{R}, X_i \quad \text{a. s.}$$

且有 $E(v_i)=0, \mathrm{var}(v_i)=1$。

注意对任意 $\alpha \in (0,1)$，由式(11.3)我们有

$$q_\alpha(X_i) = \max\{X_i'\beta + c_\alpha \sigma(X_i), 0\}$$

其中 $q_\alpha(\cdot)$ 表示给定 X_i 时，Y_i 的 α 条件分位数函数，c_α 定义为 v_i 分布（未知）的 α 分位数。于是，对于不同的 α_1 与 α_2，对任意的 $q_{\alpha_j}(X_i)>0$，我们有

$$q_{\alpha_j}(X_i) = X_i'\beta + c_{\alpha_j}\sigma(X_i), j=1,2 \tag{11.8}$$

Chen 和 Khan(2000) 提出了 β 的两个估计量，一个假定 v_i 具有已知的参数分布，另一个则未使用此假定。我们只讨论后一个估计量，由于其未假定 v_i 的分布先验已知，故能构成更一般情形。

由式(11.8)可以证明(见习题 11.2)

$$\bar{q}_\alpha(X_i) = X_i'\beta + \frac{\bar{c}}{\Delta c}\Delta q_{\alpha(X_i)}, j=1,2 \tag{11.9}$$

其中 $\bar{q}_\alpha(\cdot) = (q_{\alpha_1}(\cdot)+q_{\alpha_2}(\cdot))/2$，$\Delta q_\alpha(\cdot) = q_{\alpha_2}(\cdot) - q_{\alpha_1}(\cdot)$，$\bar{c} = (c_{\alpha_2}+c_{\alpha_1})/2$，$\Delta c = c_{\alpha_2} - c_{\alpha_1}$。式(11.9)意味着可以对 $\hat{\bar{q}}(\cdot)$ 关于 X_i 与 $\Delta\hat{q}(\cdot)$ 进行回归来估计 β（和冗余参数 $\gamma_1 = \bar{c}/\Delta c$），其中 $\hat{q}_{\alpha_j}(\cdot)$ 为 $q_{\alpha_j}(\cdot)$ 的某一非参数估计量 $(j=1,2)$。于是，Chen 和 Khan(2000) 建议通过最小化下式来估计 β（和 γ_1）：

$$\frac{1}{n}\sum_{i=1}^{n} \tau(X_i) w(\hat{q}_{\alpha_1}(X_i)) [\hat{\bar{q}}(X_i) - X_i'\beta - \gamma_1 \Delta\hat{q}_\alpha(X_i)]^2 \tag{11.10}$$

其中 $w(\cdot)$ 是一个“平滑”权函数，只对第一步中超过截断值的估计值保留观测值[①]，$\tau(\cdot)$ 是一个具有紧支撑的修剪函数。在一系列正则条件下，可以证明估计量 $\hat{\beta}$ 具有参数形式的 $\sqrt{n}$ 收敛率且具有极限正态分布。 338

Cosslett(2004) 提出了截断和断尾回归模型的（不同的）渐近有效的似然半参数估计量。他的方法基于在一个局部可观测回归中估计残差的密度函数，类似于非参数最大似然估计量，且考虑到了半参数有效得分的估计。

11.4　单变量 Kaplan-Meier 累积分布函数估计量

当出现截断数据时，还存在一类使用了所谓的 Kaplan-Meier 累积分布函数

---

① 即 $\mathbf{1}(\hat{q}_{\alpha_1}(X_i)>0)$ 的平滑形式。

估计量的半参数估计量。我们简要给出这类估计量的描述,同时也给出一个关于其应用的说明性的例子。

在许多应用中,对随机样本的所有成员都做出完备的测定通常是不现实的。例如,在医疗跟踪调查中,研究接受治疗后的存活期的分布时,某些个体可能在其死亡前就与跟踪调查失去联系。类似地,对真空管寿命的观测可能会由于测试设施移作他用而终止。以上两个例子中数据都可能被截断。

Kaplan 和 Meier(1958)提出一种新的乘积极限(product-limit)估计量用来估计累积分布函数 $F(\cdot)$ 或者生存函数 $S(\cdot)=1-F(\cdot)$。令 $Y_1,\cdots,Y_n$ 为从分布 F 或 S 抽取的感兴趣的随机样本(如病人的存活月数)。令 $L_1,\cdots,L_n$ 为一独立于所有 Y_i 的(随机或固定)截断变量。定义

$$Z_i=\min\{Y_i,L_i\},\delta_i=\mathbf{1}(Y_i\leqslant L_i) \tag{11.11}$$

然后我们观察 Z_i 与 δ_i;然而,若 $\delta_i=0$,Y_i 的准确值是未知的。在医疗跟踪调查的例子中,Y_i 是存活时间,如果在调查结束之前观测到了病人的死亡,那么 Y_i 就可观测到($\delta_i=1$)。然而,如果在时期 L_i(即还在调查进行期间)失去了与病人的联系,那么我们就不能得到 Y_i 的准确值,事实上,我们只知道 $Y_i>L_i$。另外,如果病人在调查结束后依然存活,我们也不能观察到 Y_i 的准确值。

339

定义一递增的点列 $c_0,c_1,\cdots,c_m$,累积分布函数 F(或生存函数 S)在其上取值。另定义 $I_j=\mathbf{1}(Y>c_j)$,当 Y 大于 c_j 时,取值为1,反之为0。由于 I_j 是随机变量,由条件概率公式我们有(因为 $I_j=1$ 意味着 $I_{j-1}=1$)

$$\begin{aligned}P(I_j=1\mid I_{j-1}=1)&=\frac{P(I_j=1)}{P(I_{j-1}=1)}\\&=1-\frac{P(c_{j-1}<Y\leqslant c_j)}{P(Y>c_{j-1})}\end{aligned} \tag{11.12}$$

通过取足够小的 c_1 值,如取小于数据中观测值的最小值,我们总能保证 $P(I_0=1)=1$。也就是说,所有项初期都处于存活状态。在不存在截断的情形下,我们用下式来估计式(11.12):

$$\begin{aligned}\tilde{P}(I_j=1\mid I_{j-1}=1)&=\frac{\tilde{P}(I_j=1)}{\tilde{P}(I_{j-1}=1)}=\frac{[\#Y_i>c_j]}{[\#Y_i>c_{j-1}]}\\&=1-\frac{[\#\text{ 区间 }c_{j-1}<Y_i\leqslant c_j\text{ 中的 }Y_i]}{[\#Y_i\geqslant c_{j-1}]}\end{aligned} \tag{11.13}$$

重复运用 $\tilde{P}(I_j=1)=\tilde{P}(I_j=1\mid I_{j-1}=1)\tilde{P}(I_{j-1}=1)$ 可以得到以下用来估计生存函数 $S(c_j)=P(I_j=1)$ 的估计量($P(I_0=1)=1$ 应用于下面的式(11.14)):

$$\tilde{P}(I_j = 1) = \prod_{s=2}^{j} \tilde{P}(I_s = 1 \mid I_{s-1} = 1)$$

$$= \frac{[\#Y_i > c_j]}{[\#Y_i > c_{j-1}]} \cdots \frac{[\#Y_i > c_1]}{[\#Y_i > c_0]}$$

$$= \frac{[\#Y_i > c_j]}{[\#Y_i > c_1]} = \frac{1}{n}[\#Y_i > c_j] = 1 - \hat{F}^n(c_j) \qquad (11.14)$$

其中$\hat{F}^n(c_j) = n^{-1}[\#Y_i \leqslant c_j]$为标准的经验累积分布函数。

当数据被截断时，Kaplan 和 Meier(1958)提出修改式(11.13)如下：　340

$$\hat{P}(I_j = 1 \mid I_{j-1} = 1) = 1 - \frac{[\#\text{区间 } c_{j-1} < Y_i \leqslant c_j \text{ 中的未截断 } Y_i]}{[\#Y_i \geqslant c_{j-1}]} \qquad (11.15)$$

存活率通过下式来计算：

$$\hat{S}(c_j) \equiv \hat{P}(I_j = 1) = \prod_{s=2}^{j} \hat{P}(I_s = 1 \mid I_{s-1} = 1) \qquad (11.16)$$

累积分布函数通过下式来估计：

$$\hat{F}(c_j) = 1 - \hat{S}(c_j) = 1 - \prod_{s=1}^{j} \hat{P}(I_{j-1} = 1 \mid I_{j=1}) \qquad (11.17)$$

我们借用 Kaplan 和 Meier(1958)的一个例子来说明如何应用式(11.15)和式(11.16)来估计存活率。假定初始有 100 个观察项(即病人)。定义 n_j 为存活期超出 c_{j-1}的项的数目，γ_j 为在 c_{j-1}与 c_j 期间死亡的数目，λ_j 为 c_j 时截断(丢失)的项数(假定对应于 $Y_i > c_j$)。详细的数据在表 11.1 中给出，同时给出的还有基于 Kaplan 和 Meier 的乘积极限估计量的估计以及忽略截断的传统估计量，其中我们假定截断(缺失的观测值)只出现在 1.7、3.6 和 5。

表 11.1 中 Kaplan-Meier 条件存活率$\hat{P}^{km}_{I_j \mid I_{j-1}} = \hat{P}(I_j = 1 \mid I_{j-1} = 1)$是通过式(11.15)计算出的。例如，$\hat{P}(I_2 = 1 \mid I_1 = 1) = 1 - 5/97 = 92/97$，这是由于有 97 项存活超过了 c_1，而 c_1 和 c_2 之间发生了 5 例死亡(不包括截断项)。生存函数$\hat{S}^{km}$可以使用式(11.16)作为条件概率的乘积得到。例如，$\hat{S}^{km}(c_2) = (97/100)(92/97) = 0.92$。注意传统方法得到的$\tilde{S}^n(c_j)$不考虑截断，将得到 $\hat{P}^n_{I_j \mid I_{j-1}} = 1 - 25/97 = 72/97$，这是由于其对截断项和未截断项同样对待，即均被当作死亡对待(20 + 5 = 25 例死亡)，这将导致对真实存活率的低估，即$\tilde{S}^n(c_2) = (97/100)(72/97) = 0.72 < \hat{S}^{km}(c_2)$。

表 11.1 存活的频数估计量与 Kaplan-Meier 估计量及截断存活时间数据的累积分布函数

c_{j-1}—c_j	n_j	γ_j	λ_j	$\hat{P}^{km}_{t_j \mid t_{j-1}}$	$\hat{S}^{km}(c_j)$	$\hat{P}^{n}_{t_j \mid t_{j-1}}$	$\hat{S}^{n}(c_j)$
0—1	100	3	0	97/100	0.97	97/100	0.97
1—1.7	97	5	20	92/97	0.92	72/97	0.72
1.7—2	72	4	0	68/72	0.87	68/72	0.68
2—3	68	10	0	58/68	0.74	58/68	0.58
3—3.6	58	9	12	49/58	0.63	37/58	0.37
3.6—4	37	6	0	31/37	0.52	31/37	0.31
4—5	31	15	16	16/31	0.27	0/31	0

注:其中上标 km 表示 Kaplan-Meier,n 表示忽略截断的传统估计量。

341 Kaplan-Meier 乘积极限估计量已应用于多种具有截断数据的非参数与半参数模型中。例如,Horowitz(1986,1988)应用 Kaplan 和 Meier(1958)的方法构建了一个具有截断数据的线性回归模型的估计量。在讨论估计具有截断数据的非参数和半参数模型之前,我们先来讨论截断数据的多变量累积分布函数估计。

11.5 多变量 Kaplan-Meier 累积分布函数估计量

下述讨论中,我们保留前一节中使用的符号,即 $Z_i = \min\{Y_i, L_i\}$,$\delta_i = \mathbf{1}(Y_i \leq L_i)$。令 $Z_{i:n}$ 表示有序的满足 $Z_{1:n} \leq Z_{2:n} \leq \cdots \leq Z_{n:n}$ 的 Z 值,$\delta_{[i:n]}$ 表示与 $Z_{i:n}$ 相对应的 δ 值。有了这些符号,上节最后讨论的 $F(y)$ 的 Kaplan-Meier 估计量即如下乘积极限估计量 $\hat{F}_n(y)$:

$$1 - \hat{F}_n(y) = \prod_{i=1}^{n} \left[1 - \frac{\delta_{[i:n]}}{n-i+1}\right]^{\mathbf{1}(Z_{i:n} \leq y)} \tag{11.18}$$

Kaplan 和 Meier(1958) 证明了式(11.18)定义的 $\hat{F}_n(\cdot)$ 是 $F(\cdot)$ 的最大似然估计量。

342 Stute(1993) 把乘积极限估计量扩展到了多变量框架。令 $F(y,x)$ 表示 (Y,X) 的联合累积分布函数,其中 X 是一个 q 维连续随机变量。Stute 提出用下式来估计 $F(y,x)$:

$$\hat{F}_n(y,x) = \sum_{i=1}^{n} W_{in} \mathbf{1}(X_{[i:n]} \leq x, Z_{i:n} \leq y) \tag{11.19}$$

其中

$$W_{in} = \frac{\delta_{[i:n]}}{n-i+1} \prod_{j=1}^{i-1} \left[\frac{n-j}{n-j+1}\right]^{\delta_{[j:n]}} \tag{11.20}$$

是在 Kaplan-Meier 估计量中关于 $Z_{i:n}$ 的质量(mass),$X_{[i:n]}$ 为对应于 $Z_{i:n}$ 的 X 的观测值。

Stute(1993)证明了$\hat{F}_n(y,x)$的一致性。事实上,Stute 证明了对一大类统计函数 S,有 $S(\hat{F}_n) \to S(F)$ 依概率 1 收敛。Stute(1996)进一步建立了$\hat{F}_n(\cdot)$的渐近分布,具体是通过考虑一个一般的线性函数$\hat{F}_n$,定义为

$$S_n^{\phi} \stackrel{\text{def}}{=} \sum_{i=1}^{n} W_{in}\phi(X_{[i:n]}, Z_{i:n}) \equiv \int \phi \mathrm{d}\hat{F}_n \tag{11.21}$$

其中$\hat{F}_n$ 为定义在式(11.19)中的 $F(y,x)$ 的 Kaplan-Meier 估计量,$\phi(\cdot,\cdot)$为$(X_{[i:n]}, Z_{i:n})$的一个平滑函数。

在一些正则条件下,Stute(1996) 证明了

$$\sqrt{n}(S_n^{\phi} - S^{\phi}) \to N(0, \sigma_{\phi}^2) \text{依分布收敛} \tag{11.22}$$

其中 σ_{ϕ}^2 是$\sqrt{n}S_n^{\phi}$ 的渐近方差(σ_{ϕ}^2 的显性表达式见 Stute(1996,p.464))。

注意从式(11.19)中移去 x 可得到一个 $F(y)$的不同估计量:

$$\tilde{F}_n(y) = \sum_{i=1}^{n} W_{in}\mathbf{1}(Z_{i:n} \leqslant y) \tag{11.23}$$

其中 W_{in} 在式(11.20)中定义。习题 11.3 要求读者证明式(11.18)中定义的$\hat{F}_n(y)$与式(11.23)中定义的$\tilde{F}_n(y)$确实是相同的。

Cai(2001)把 Kaplan 和 Meier(1958)的乘积极限估计量推广到了相依数据(dependent data)情形,同时建立了对 α 混合数据(α-mixing data)估计量的强一致性。Lai 和 Ying(1991)以及 Lin 和 Ying (1999)也探讨了存在截断数据时分布函数的非参数估计问题。 343

11.5.1 具有随机截断的非参数回归模型

接下来,我们讨论一个具有随机截断相依数据的非参数回归模型的估计。我们重点关注 Fan 和 Gijbels(1992)以及 Cai(2003)提出的局部线性估计量。模型如下:

$$Y_i = g(X_i) + u_i \tag{11.24}$$

其中 Y_i 为因变量,可以是一个个体的存活时间或者某些试验研究的对象,X_i 是相对应的 q 维协变量。为了表述方便,我们假定 X_i 是一个 q 维连续随机变量向量。截断方法下,不是观察 Y_i,而是观察 $Z_i = \min(Y_i, L_i)$ 和 $\delta_i = \mathbf{1}(Y_i \leqslant L_i)$。因此,观测值是来自$\{Z,X,\delta\}$的一个随机样本$\{Z_i, X_i, \delta_i\}_{i=1}^{n}$。

首先,注意不存在截断数据时,$g(\cdot)$的传统的局部线性估计量可以通过关于 a 和 b 最小化以下目标函数(a 估计的是 $g(x_0)$)来得到:

$$\int [y - a - (x - x_0)'b]^2 K_{h,x,x_0} \mathrm{d}F_n(x,y)$$

$$= \frac{1}{n}\sum_{i=1}^{n}[Y_i - a - (X_i - x_0)'b]^2 K_{h,X_i,x_0} \qquad (11.25)$$

其中 $K_{h,x,x_0} = K_h(x,x_0)$是乘积核函数,$F_n(x,y)$是不存在截断时$\{X_i,Y_i\}_{i=1}^n$的(标准)经验分布。存在截断数据时,Fan 和 Gijbels(1992)以及 Cai(2003)建议用 $\hat{F}_n(\cdot,\cdot)$来替换式(11.25)中的经验累积分布函数 $F_n(\cdot,\cdot)$,$\hat{F}_n(\cdot,\cdot)$是 $F(\cdot,\cdot)$的 Kaplan 和 Meier(1958)估计量。于是,我们通过最小化以下目标函数来估计 $g(x_0)$:

$$\int [y - a - (x - x_0)'b]^2 K_h(x,x_0)\,\mathrm{d}\,\hat{F}_n(x,y)$$

$$= \sum_{i=1}^{n} W_{in}[Z_{i;n} - a - (X_{[i;n]} - x_0)'b]^2 K_{h,X_{[i;n]},x_0} \qquad (11.26)$$

344 其中 W_{in},$Z_{i;n}$和 $X_{[i;n]}$均按 11.5 节中的定义。

令$\hat{a}(x_0)$和$\hat{b}(x_0)$表示最小化式(11.26)时 a 与 b 的值,那么$\hat{a}(x_0)$估计了 $g(x_0)$。由标准最小二乘公式,我们有(用 x 替换 x_0)

$$\begin{bmatrix}\hat{a}(x)\\ \hat{b}(x)\end{bmatrix} = \left[\sum_{i=1}^{n} W_{in}K_h(X_{[i;n]},x)[1,(X_{[i;n]}-x)']'[1,(X_{[i;n]}-x)']\right]^{-1}$$

$$\times\left[\sum_{i=1}^{n} W_{in}K_h(X_{[i;n]},x)[1,(X_{[i;n]}-x)']'Z_{i;n}\right] \qquad (11.27)$$

Fan 和 Gijbels(1992)以及 Cai(2003)证明了如下结果,其中 $\hat{g}(x)=\hat{a}(x)$如式(11.27)中所定义。

定理 11.1 在一定的正则条件下,有

$$\sqrt{nh_1\cdots h_q}\left[\hat{g}(x) - g(x) - \frac{\kappa_2}{2}\sum_{s=1}^{q} h_s^2 g_{ss}(x)\right]\xrightarrow{d} N(0,\Sigma_x)$$

其中 $g_{ss}(x) = \partial^2 g(x)/\partial x_s^2$,$\Sigma_x = \kappa^q v(x)/f(x)$,其中 $v(x) = E[Y^2\{1 - G(Y)\}^{-1}]$,$G(\cdot)$为 L_i 的累积分布函数(即截断变量)。

注意渐近偏误与未截断数据的局部线性估计量所得到的是相同的,这是很有趣的现象。然而,截断情形下的渐近方差是大于相应未截断情形下的渐近方差的。当完全不存在截断时,可以设定 $G=0$,则 Σ_x 退化为 $\kappa^q E(Y_i^2|X_i=x)/f(x)$,正是未截断情形下的渐近方差。

Cai(2003)进一步证明了$\lim_{n\to\infty}\mathrm{var}(\xi_{n,i}(x)) = \kappa^q v(x)f(x)$,其中 $\xi_{n,i}(x) =$

$\sqrt{h_1 \cdots h_q} K_h(X_i,x)[1-G(Z_i)]^{-1}\delta_i$，中间涉及了未知的截断分布 $G(\cdot)$。在变量 L_i 不依赖协于变量 X_i 的假定下，可以使用 Kaplan-Meier 乘积极限估计量来估计 $G(\cdot)$，即 $1-\hat{G}(x) = \prod_{j, Z_{j:n} \le z} \{(n-j)/(n-j+1)\}^{1-\delta_{[j:n]}}$。这可以导出 $\xi_{n,i}(x)$ 的一个估计量 $\hat{\xi}_{n,i}(x) = \sqrt{h_1 \cdots h_q} K_h(X_i,x)[1-\hat{G}(Z_i)]^{-1}\delta_i$。因此，$\Sigma_x$ 的一个一致估计量为 $n^{-1}\sum_{i=1}^{n}\{\hat{\xi}_{n,i}(x)-\bar{\xi}_n(x)\}^2\delta_i$，其中 $\bar{\xi}_n(x) = n^{-1}\sum_{i=1}^{n}\hat{\xi}_{n,i}(x)$。

其他的相关工作包括 Fan, Gijbels 和 King(1997)，他们提出了一个局部似然和半参数局部似然方法，用来估计半参数风险函数，以及 Cai 和 Sun (2003)，他们应用了局部似然方法来估计半参数变系数风险函数。

11.6　非参数截断回归

345

11.6.1　Lewbel 和 Linton 的方法

Lewbel 和 Linton(2002)考虑了形如 $Y_i = \max\{g(X_i)-\varepsilon_i, c\}$ 的截断回归模型，其中 $g(\cdot)$ 是对未截断总体的条件期望，截断点 c 假定为一已知常数(模型写作 $g-\varepsilon$ 而非通常的 $g+\varepsilon$ 是因为可以简化后面的结果)。若有 $E(\varepsilon_i)=0$，则函数 $g(X_i)$ 等同于未截断总体的回归函数。

Lewbel 和 Linton(2002) 考虑了函数 $g(\cdot)$ 和 $F(\cdot)$ (ε_i 的分布)的完全非参数估计，这两个函数都假定是未知的。给定条件下，$g(\cdot)$ 可以被一致估计，F 在给定 $g(\cdot)$ 时也可被估计。

以下讨论中，我们定义如下函数：$r(x) = E(Y|x)$，$q(r) = E[\mathbf{1}(Y>0)|r(X)=r]$。Lewbel 和 Linton(2002) 证明了对某些较大的常数 λ_0，

$$g(x) = \lambda_0 - \int_{r(x)}^{\lambda_0} \frac{1}{q(r)} \mathrm{d}r \tag{11.28}$$

式(11.28)提供了估计 $g(\cdot)$ 的基础。Lewbel 和 Linton (2002) 提出了下面的两步过程。首先，使用一个局部多项式估计量来估计回归函数 $E(Y_i|X_i) = r(X_i)$，称其为 $\hat{r}(X_i)$。接下来，令 $\hat{q}(r)$ 为 $\mathbf{1}(Y_i>0)$ 关于生成的回归元 $\hat{r}(X_i)$ 的一维非参数回归。那么，对某些固定的正 λ_0，令

$$\hat{g}(X_i) = \lambda_0 - \int_{\hat{r}(X_i)}^{\lambda_0} \frac{1}{\hat{q}(r)} \mathrm{d}r$$

此单变量积分易通过数值方法来计算；于是也可以方便地计算出 $\hat{g}(X_i)$。令 $\lambda_r = \sup_{x\in\Omega} r(x)$，$\lambda_\varepsilon = \sup_{\varepsilon\in\Omega_\varepsilon}\varepsilon$，其中 Ω 和 Ω_ε 分别是 x 和 ε 的支撑。若有 $\lambda_r >$

λ_g，在实际应用中，可以用 $\hat{\lambda} = \max_{i=1,\cdots,n} \hat{r}(X_i)$ 或者 $\hat{\lambda}_Y = \max_{i=1,\cdots,n} Y_i$ 来代替 λ_0。

令 $r_l(x) = \partial r(x)/\partial x_l$ ($l = 1,\cdots,q$), $s(x) = E[\mathbf{1}(Y_i^* > 0) | X_i = x]$，偏导数估计为

$$\hat{g}_l(x) = \frac{\hat{r}_l(x)}{\hat{s}(x)}$$

其中 $\hat{r}_l(x)$ 与 $\hat{s}(x)$ 分别为 $r_l(x)$ 与 $s(x)$ 的局部多项式估计量。

346 Lewbel 和 Linton (2002)给出了 $\hat{g}(x)$ 和 $\hat{g}_l(x)$ 能够分别一致且渐近正态估计 $g(x)$ 与 $g_l(x)$ 的条件。特别地，令 $\sigma_r^2(x) = \mathrm{var}(Y | X = x)$，$f(x)$ 为 X 的概率密度函数，Lewbel 和 Linton 证明了

$$\sqrt{nh_1\cdots h_q}\Big(\hat{g}(x) - g(x) - \sum_{s=1}^{q} h_s^2 b_s(x)\Big) \xrightarrow{d} N\Big(0, \frac{\kappa^q \sigma^2(x)}{f(x)s^2(x)}\Big) \quad (11.29)$$

其中 $\kappa = \int k(v)^2 \mathrm{d}v$，$k(\cdot)$ 是局部线性核估计量中乘积核所使用的单变量核函数，$b_s(\cdot)$ 是一些平滑且有界的函数。平滑参数需要满足的条件是 $h_l \to 0$（所有 $l = 1,\cdots,q$ ）且 $(nh_1\cdots h_q)\sum_{s=1}^{q} h_s^4 < \infty$。如果将平滑条件替换为一个弱平滑的条件 $(nh_1\cdots h_q)\sum_{s=1}^{q} h_s^4 \to 0$，那么偏误首项可以被忽略掉且估计量 $\hat{g}(x) - g(x)$ 变为真正的以零为中心。

11.6.2 Chen, Dahl 和 Khan 的方法

Chen, Dahl 和 Khan (2005)对于非参数截断回归模型提出了一个替代的估计方法。为了解决对小于截断点范围内位置函数的估计感兴趣带来的一系列问题，Chan, Dahl 和 Khan 提出了一个非参数位置—尺度(location-scale)模型的扩展来处理数据。他们考虑模型

$$Y_i^* = g(X_i) + \sigma(X_i)\varepsilon_i$$
$$Y_i = \max\{Y_i^*, 0\}$$

其中 Y_i^* 是未观测的潜因变量，Y_i 是可观测的因变量，当超出固定的截断点(零点)时等于 Y_i^*，其他情形则为 0。X_i 是可观测的 q 维随机向量，ε_i 是具有零均值、单位方差的独立同分布扰动项，其分布独立于 X_i。

Chen, Dahl 和 Khan (2005)给出了在施加了位置约束后，$g(x)$ 可以被识别以及估计所需要的条件，即 ε_i 的中位数为零，他们可以在 X_i 的整个支撑集上识别 $g(x)$，而不是仅仅在超出截断点的范围可以识别。他们的方法基于条件均

值与满足 $g(X_i) \geqslant 0$ 的 X_i 的上分位点之间的结构关系，且要求满足条件 $P(g(X) \geqslant 0) > 0$ 。

在 ε_i 与 X_i 独立的假定下，也假定 ε_i 的概率密度函数在 $\mathbb{R}$ 上为正，这时，存在分位点 $\alpha_1 < \alpha_2 < 1$ 使得 347

$$q_{\alpha_j}(x) = g(x) + c_{\alpha_j}\sigma(x) > 0, \quad j = 1,2 \tag{11.30}$$

其中 $q_\alpha(x)$ 是 Y_i 的 α 条件分位点，条件于 $X_i = x$ ，c_α 则是 ε_i 的第 α 分位值。

式(11.30)给出了估计 $g(\cdot)$ 与 $\sigma(\cdot)$ 的回归框架。一旦我们得到 $q_\alpha(\cdot)$ 和 $c_\alpha(\cdot)$ 的估计量，如 $\hat{q}_\alpha(\cdot)$ 和 $\hat{c}_\alpha(\cdot)$ ，我们就可以关于 $\hat{c}_\alpha(\cdot)$ 回归 $\hat{q}_\alpha(\cdot)$ 来得到 $q_\alpha(\cdot)$ 和 $c_\alpha(\cdot)$ 的估计。

注意从式(11.30)与 $q_{0.5}(x) = g(x)$ (由于 ε_i 具有中位数零)的事实可以导出 $q_{\alpha_j}(x) = q_{0.5}(x) + c_{\alpha_j}\sigma(x)$ ($j = 1,2$)，从这两个方程中解出 c_{α_2} 得到

$$c_{\alpha_2} = \frac{q_{\alpha_2}(x) - q_{0.5}(x)}{q_{\alpha_1}(x) - q_{0.5}(x)} \tag{11.31}$$

我们现在来讨论 Chen，Dahl 和 Khan(2005) 给出的三步估计过程。

(i) 使用一个局部常数方法对每个观察点 X_i 估计 $q_{0.5}(X_i)$ 。令 $\bar{q}_{0.5}(X_i)$ 表示用来决定那些中位数超出截断值 c 的观测值的估计值。

(ii) 计算局部多项式估计量的一个加权平均来估计未知扰动分位数 c_{α_l} ，$l = 1,\cdots,N$ ，其中 $\hat{c}_{\alpha_l}$ 等于

$$\frac{1}{n}\sum_{i=1}^{n}\tau(X_i)w(\bar{q}_{0.5}(X_i))\frac{(\tilde{q}_{\alpha_l}(X_i) - \tilde{q}_{0.5}(X_i))/(\tilde{q}_{\alpha_s}(X_i) - \tilde{q}_{\alpha_l}(X_i))}{\frac{1}{n}\sum_{j=1}^{n}\tau(X_j)w(\bar{q}_{0.5}(X_j))}$$

其中 $w(\cdot)$ 是一权重函数，符号“～”表示第二步估计量(这是为了和第一步估计量(符号为 －) $\bar{q}_{0.5}(x)$ 进行区别)。

(iii) 使用局部多项式方法，在感兴趣的点 x (x 可以不是观测点)处估计 $q_\alpha(\cdot)$ ，即

$$q_{\alpha_l}(x) = g(x) + c_{\alpha_l}\sigma(x)$$

对任意满足 $q_{\alpha_l}(x > 0)$ 的 α_l 均成立。令 $\theta(x) = (g(x),\sigma(x))'$ ，$\hat{\theta}(x) = (\hat{g}(x),\hat{\sigma}(x))'$ 。令 $\hat{\boldsymbol{c}}_{\alpha_l} = (1,\hat{c}_{\alpha_l})$ ，$\tilde{\hat{d}}_{\alpha_l}$ 表示示性函数 $\mathbf{1}[\hat{q}_{\alpha_l} \geqslant \varepsilon]$ ，其中 $\varepsilon > 0$ 为一较小的常数，估计量 $\theta(x)$ 为 348

$$\hat{\theta}(x) = \left(\sum_{l=1}^{N}\tilde{\hat{d}}_{\alpha_l}\hat{\boldsymbol{c}}_{\alpha_l}\hat{\boldsymbol{c}}'_{\alpha_l}\right)^{-1}\sum_{l=1}^{N}\tilde{\hat{d}}_{\alpha_l}\hat{\boldsymbol{c}}_{\alpha_l}\hat{q}_{\alpha_l}(x) \tag{11.32}$$

其中 $\hat{g}_\alpha(x)$ 为第三步中 $g_\alpha(x)$ 的非参数估计量。

令 h_{js} ，$j = 1,2,3$ ，表示第 j 步中关于 X_{is} ($s = 1,\cdots,q$)的平滑参数，Chen，

Dahl 和 Khan(2005)假定,对所有的 $s = 1,\cdots,q$,

(i) $h_{1s} \to 0$ 且 $nh_{11}\cdots h_{1q}/\ln(n) \to \infty$ 。

(ii) $h_{2s} = o(n^{-1/(2p+q)})$ 且 $nh_{21}\cdots h_{2q}/\ln(n) \to \infty$ 。

(iii) 对某一有限的正的常数 c_s 有 $h_{3s} = c_s n^{-1/(2p+q)}$ 。

注意第一步估计量仅在计算 $\hat{c}_{\alpha_l}$ 时出现在 $w(\bar{q}_{0.5}(X_i))$ (使用 h_{1s})中,其中 $w(\cdot)$ 用来选择中位数为正的观测值。第二步估计量为 $\tilde{q}_{\alpha}(X_i)$ (使用 h_{2s}),也被用来计算 $\hat{c}_{\alpha_l}$ 。最后,第三步估计量为 $\hat{g}(x)$ (使用 h_{3s}),以其来计算式(11.32)中定义的 $\hat{\theta}(x)$ 。

估计量的渐近性质参考 Chen,Dahl 和 Khan(2005) 。

11.7 习题

习题 11.1 推导式(11.4)。

习题 11.2 推导式(11.9)。

习题 11.3 证明式(11.18)中定义的 $\hat{F}_n(y)$ 与式(11.23)中定义的 $\tilde{F}_n(y)$ 相同。

习题 11.4 推导式(11.31)。

第 3 部分
一致模型的设定检验

第12章 模型设定检验

如同我们在第1章开始讨论的,正确设定参数模型是估计和推断问题的最佳解决办法。但遗憾的是,理论在恰当的函数设定方面很少提供任何指导。因此,在现实中我们必然会面临参数模型可能设定不正确的情形,这就典型地形成"模型设定检验"。很多常用的参数模型设定检验本质上要求用户设定对立参数集,如果数据生成过程服从设定的对立模型,原假定将被拒绝。如果存在一些这个检验不能检测的对立模型,那么这个检验被称为一个"非一致检验",因为它缺少某一方面的检验功效。非参数方法一个普遍的应用是检测不正确设定的参数模型。我们现在将讨论如何使用非参数技术来构造一致的模型设定检验。 351

我们首先给出"一致检验"的定义。定义 H_0 为一个原假定,我们希望检验此原假定的有效性。如果当 $n\to\infty$,P(拒绝 $H_0|H_0$ 不成立)$\to 1$,那么一个检验被称为一个一致性检验。

检验功效定义为 P(拒绝 $H_0|H_0$ 不成立)。因此,一个一致性检验具有等于1的渐近功效。假定我们要检验两个地区的收入分布(或者可能是同一个地区在两个不同时期的分布)是否相同。用 $f(\cdot)$ 和 $g(\cdot)$ 分别表示这两个收入分布。那么对于所有的 $x\in\mathbb{R}$,原假定为 $f(x)=g(x)$。参数方法以为 $f(\cdot)$ 和 $g(\cdot)$ 假定一个参数族开始,例如,假定 $f(\cdot)$ 和 $g(\cdot)$ 都属于正态分布族。因为一个正态分布可以完全由它的均值和方差刻画,我们将对这两个分布的均值和方差是否相同进行检验,如果正态性假定是正确的,那么当然这个参数方法有效。然而,当潜在分布设定不正确时,参数检验将得到误导性的结果。例如,$f(\cdot)$ 和 $g(\cdot)$ 可能不属于相同的分布族,尽管这两个分布可能事实上具有相同的均值和方差。在这样的情况下,基于正态性假定的参数检验 352

可能会不拒绝一个不正确的原假定,也就是没有检验功效。因此,参数检验不是一个一致性检验。在这个例子中,当正态性假定正确时,参数检验具有等于1的渐近检验功效,但是当假定的分布设定不正确时,则可能没有检验的功效。也就是说,对每一个偏离原假定的可能,它将不具有等于1的渐近功效。

与对应的参数方法不同,非参数方法可以用来构造在每一个方面都具有检验功效的一致性检验。我们可以采用非参数技术来估计 $f(\cdot)$ 和 $g(\cdot)$,然后检验对于全部的 x 是否都有 $f(x)=g(x)$。$f(\cdot)$ 和 $g(\cdot)$ 这两个分布之间一个方便的距离测度是它们的积分平方差:

$$J = \int [f(x) - g(x)]^2 \mathrm{d}x$$

我们可以观察到当 H_0 成立时,$J=0$;H_0 不成立时,$J>0$。因此,J 值为一致检验 H_0 提供了一个良好的基准。我们将在12.2节讨论如何基于一个 J 的核估计来检验 H_0。

让我们考虑另一个例子,在这个例子中我们的目的是检验以下形式的线性回归模型是否设定正确:

$$Y_i = \alpha + X_i'\beta + u_i \tag{12.1}$$

其中,X_i 是 $q\times 1$ 维的变量,$\beta=(\beta_1,\cdots,\beta_q)'$,$\alpha$(一个标量)是未知参数。

通过把式(12.1)嵌套进一个更大的参数模型,传统的为检验遗漏变量的 F 检验和 t 检验可以用于检验式(12.1)的形式是否正确。但是,这些基于参数模型的检验只在某些方面具有检验功效,正如我们下面要讨论的。

353 为了表达简洁,我们假定 X_i 是一个标量。标准 t 检验(或者 F 检验)可以用于检验假定

$$H_{p0}: Y_i = \alpha + \beta_1 X_i + u_i \tag{12.2}$$

备择假定为

$$H_{p1}: Y_i = \alpha + \beta_1 X_i + \beta_2 X_i^2 + u_i \tag{12.3}$$

如果 H_{p1} 是真实模型,则一个参数 t 或者 F 检验将具有功效来检验真实模型是 H_{p1} 的事实。因此,当对原假定模型的偏离落在 H_{p1} 的方向内时,一个参数检验被期望具有较好的检验功效。

但是,我们注意到 H_{p1} 并不是 H_{p0} 的补集。也就是说,当 H_{p0} 不成立时,并不必然意味着 H_{p1} 是正确的。事实上,假定下面的方程代表真实模型:

$$H_{p2}: Y_i = \alpha + \beta_1 X_i + \beta_3 X_i^3 + u_i \tag{12.4}$$

如果 H_{p2} 是真实的模型,则 H_{p0} 或者 H_{p1} 都不是正确设定的。现在我们进一步假定 X_i 是以0为中心的对称分布(因此,对于 $l=1,3,5$,有 $E(X_i^l)=0$)。这时,$\hat{\beta}_2 \xrightarrow{p} 0$(见习题12.1),这里 $\hat{\beta}_2$ 是基于模型(12.3)对 β_2 的最小二乘估计值。

这意味着基于模型(12.3)估计的一个标准 t 检验(或者 F 检验)将缺乏检验功效,在实践中很可能没有拒绝错误的原假定模型(12.2)。因此,当原假定不成立时,一个参数检验不能在每一个方向都具有检验功效,从而不能构成一致检验的基础。然而非参数方法可以用来构造一个在所有方向都偏离原假定模型,但仍具有检验功效的模型设定检验,这解释了为什么非参数方法如此具有吸引力。Bierens(1982)应该是第一个提出了对正确参数回归函数形式的一致(非平滑)检验,而 Ullah(1985)使用了非参数平滑技术来考虑一致模型设定检验。

在这一章中余下的部分我们将讨论如何使用非参数方法来构造一致的核估计模型设定检验。也可以不使用非参数技术来构造模型设定检验,我们把这第二种类型的检验称为"非平滑检验"(我们将在第13章讨论这类检验)。一致 k 近邻估计和序列模型设定检验将分别在第14和第15章中讲述。

12.1 一个简单的关于参数回归模型函数形式的一致检验

354

我们可以基于 $\int[\hat{g}(x)-\hat{\alpha}-x'\hat{\beta}]^2$ 来检验一个线性回归模型是否正确设定,其中 $\hat{g}(x)$ 是 $g(x)=E(Y|x)$ 的非参数估计,$\hat{\alpha}$ 和 $\hat{\beta}$ 是线性回归模型 α 和 β 的最小二乘估计值。但是,如我们即将证明的,使用这种基于回归的检验,以一个条件矩检验的形式,它的原假定是 $H_0^a:E(u|x)=0$,对几乎所有的 x 都是可取的,其中 u 和 x 的定义依赖于特定的模型和要检验的原假定。在上述考虑检验一个线性回归模型的例子中,可以定义 $u=Y-\alpha-X'\beta$;因此几乎处处有 $E(u|x)=0$,相当于几乎处处有 $E(Y|x)=\alpha+x'\beta$。

我们现在考虑根据 $I=E[uE(u|X)M(X)]$ 构造的一个条件矩检验,其中 $M(\cdot)$ 是一个非负加权函数。采用这个统计量,是因为根据迭代期望法则,可以观察到有 $I=E\{[E(u|X)]^2M(X)\}\geq 0$ 成立,当且仅当 H_0^a 为真时等号成立。因此,I 可以作为一致地检验 H_0^a 的一个合适的候选。常用的加权函数有:(i) $M(x)=f(x)$ ($f(\cdot)$ 是 X 的概率密度函数)。(ii) $M(x)\equiv 1$。选择 $M(x)=f(x)$ 是为了避免一个随机的分母问题,这简化了一致性分析。因此,为了简化理论分析,我们主要关注情形(i),并使用 $M(x)=f(x)$。

使用 $M(\cdot)=f(\cdot)E[uE(u|X)f(X)]$ 的样本表达式由 $n^{-1}\sum_{i=1}^{n}u_iE(u_i\mid X_i)f(X_i)$ 给出。把 u_i 和 $E(u_i|X_i)f(X_i)$ 用其一致估计量分别取代,可以得到一个可行统计量。这样得到的检验统计量与基于 $E\{[E(u|x)]^2f(X)\}$ 的样本表达式相比有一个更简单的形式,因为后者的样本表达式是 $n^{-1}\sum_{i=1}^{n}[E(u_i\mid X_i)]^2f(X_i)$。

当以非参数核估计量取代条件均值函数时，这个检验统计量将包含三个加总，而前者只包含两个加总。前一个检验统计也比基于参数和非参数估计值之间的积分平方差的检验要简单（详见12.1.2节）。

355

12.1.1 正确参数函数形式的一致性检验

为了检验一个参数回归模型的正确性，原假定为

$$H_0^a: E(Y \mid x) \equiv g(x) = m(x, \gamma_0), \text{所有 } x \text{ 和某些 } \gamma_0 \in \mathcal{B} \subset \mathbb{R}^p \quad (12.5)$$

其中 $m(x,\gamma)$ 是一个已知的函数，γ 是一个未知参数（明显地作为一种特例包含在一个线性回归模型中）的 $p \times 1$ 向量，$\mathcal{B}$ 是 $\mathbb{R}^p$ 的一个紧致子集。备择假定是 H_0^a 的反面情形，例如对于在一个正测度集（关于 x）上所有的 $\gamma \in \mathcal{B}$，$H_1^a: E(Y|x) \equiv g(x) \neq m(x,\gamma)$。如果我们定义 $u_i = Y_i - m(X_i, \gamma_0)$，那么对于所有的 x，原假定可以等价地写成

$$E(u_i \mid X_i = x) = 0 \quad (12.6)$$

我们可以构建一个基于式(12.6)的非参数核估计的一致的模型设定检验。

利用式(12.6)和在本章开篇的讨论，可以建立一个基于 $E\{u_i E(u_i | X_i) f(X_i)\}$ 的一致性检验的统计量，其中 $f(x)$ 是 X 的联合密度函数。这里简单地应用密度加权是为了避免随机分母会出现在核估计量中。

$E[u_i E(u_i|X_i) f(X_i)]$ 的样本表达式由公式 $n^{-1}\sum_{i=1}^{n} u_i E(u_i \mid X_i) f(X_i)$ 给出。为了得到一个可行的检验统计量，我们用 $\hat{u}_i$ 替代 u_i，其中 $\hat{u}_i = Y_i - m(X_i, \hat{\gamma})$ 是一个从参数原模型中得到的残差，$\hat{\gamma}_i$ 是一个基于原模型的 $\sqrt{n}$ 一致性估计量（即 γ 的非线性最小二乘估计量）。我们通过去一核估计量 $(n-1)^{-1}\sum_{j\neq i}^{n} \hat{u}_j K_{h,ij}$ 来估计 $E(u_i \mid X_i) f(X_i)$，其中 $K_{h,ij} = \prod_{s=1}^{q} h_s^{-1} k((X_{is} - X_{js})/h_s)$ 且我们假定 X_i 是一个连续的 q 维变量。我们将会在12.1.2节中讨论 X_i 是一个混合了离散变量和连续变量的向量时的情形。我们的检验统计量是基于下式

$$I_n^a \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \hat{u}_j \left\{ \frac{1}{n-1}\sum_{j=1, j\neq i}^{n} \hat{u}_j K_{h,ij} \right\} = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n} \hat{u}_i \hat{u}_j K_{h,ij} \quad (12.7)$$

356

条件 12.1

(i) $f(x)$，$g(x) = E(Y|x)$，$\sigma^2(x) = E(u_i^2 | x)$ 和 $\mu_4(x) = E(u_i^4 | x)$ 都是有界的连续函数。$m(x,\gamma)$ 对于 x 是连续的且关于 γ 二阶可微，由于函数具有有限的二阶矩，因此 $m(\cdot)$ 和其导函数都是有界的。

(ii) $k(\cdot)$ 是一个有界的二阶核，当 $n \to \infty$ 时，$h_s \to 0$ $(s = 1, \cdots, q)$ 且 $nh_1 \cdots h_q \to \infty$。

I_n 的渐近分布由下述定理给出。

定理 12.1　假定条件 12.1 成立，那么在 H_0^a 下，

$$T_n^a \stackrel{\text{def}}{=} n(h_1 \cdots h_q)^{1/2} I_n^a / \sqrt{\hat{\sigma}_a^2} \xrightarrow{d} N(0,1)$$

其中 $\hat{\sigma}_a^2 = \dfrac{2h_1 \cdots h_q}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{u}_i^2 \hat{u}_j^2 K_{h,ij}^2$ 是 $\sigma_a^2 = 2\kappa^q E(\sigma^4(X)f(X))$ 的一个一致估计量。

定理 12.1 的证明在 12.6 节中给出。下一个定理证明 T_n^a 是一个一致检验。

定理 12.2　假定条件 12.1 成立。如果 H_0^a 不成立，对于任何的正的常数 $C>0$，我们有

$$P(T_n^a > C) \to 1, \text{当 } n \to \infty$$

定理 12.2 的证明在 12.6.2 节当中概述。T_n^a 统计量由 Li(1994) 和 Zheng(1996) 独立地提出，Fan 和 Li(1996) 中也使用了它的一个版本。

注意因为 T_n^a 是单边检验，实际上如果 $T_n^a > Z_\alpha$，则 H_0^a 在显著水平 α 被拒绝，其中 Z_α 是标准正态变量的上 α 分位点。例如，如果 $T_n^a > 1.645$，则在 5% 的显著水平上拒绝 H_0^a。

定理 12.1 和定理 12.2 对 h_s 的允许范围仅有最低限度的要求。所以 I_n^a 允许平滑参数值有一个较宽的范围。例如，它允许使用最小二乘交错鉴定法、插入法和一些特别的方法来选择 h_s。用前面提到的任何一种方法来代替未知的最优非随机平滑参数时，定理 12.1 和定理 12.2 仍旧是有效的，只要它们能够使 $h_s \to 0$ 且当 $n \to \infty$ 时，$nh_1 \cdots h_q \to \infty$。　357

在原假定 H_0^a 下，可以证明 $n(h_1 \cdots h_q)^{1/2} I_n^a$ 以一个缓慢的速度 $O_p((h_1 \cdots h_q)^{1/2})$ 收敛于零均值的正态随机变量（见 12.6.1 节的式(12.41)）。如果我们考虑 $q=1$ 和 $h \sim n^{-1/5}$（在 $q=1$ 的情形下非参数估计的最优速度）的情形，则 $O(h^{1/2}) = O(n^{-1/10})$ 以一个极慢的速度收敛于零。在 Li 和 Wang(1998) 的模拟研究中确实表明 I_n^a 检验以一个非常慢的速度收敛于一个正态变量。即使样本容量 $n = 1\,000$，该检验仍显著地容量不足（undersized）。尽管如此，他们证明它是一个原始自举程序（Liu(1988)），可以用来大大改进 I_n^a 检验在有限样本中的表现（关于原始自举法工作的更深见解见问题 12.2）。

计算原始自举统计量包括的步骤如下所示：

(i) 对于 $i=1,\cdots,n$，产生一个两点原始自举误差：以概率 $r=(1+\sqrt{5})/(2\sqrt{5})$ 有 $u_i^* = [(1-\sqrt{5})/2]\hat{u}_i$，以概率 $1-r$ 有 $u_i^* = [(1+\sqrt{5})/2]\hat{u}_i$。

(ii) 得到 $Y_i^* = m(X_i, \hat{\gamma}) + \hat{u}_i^*$ $(i=1,\cdots,n)$。产生的样本 $\{X_i, Y_i^*\}_{i=1}^n$ 被称为自举样本。然后得到自举残差 $\hat{u}_i^* = Y_i^* - m(X_i, \hat{\gamma}^*)$ $(i=1,\cdots,n)$，其中 $\hat{\gamma}^*$ 是

关于自举样本 γ 的非线性最小二乘估计量。

(iii) 利用自举残差来计算检验统计量 $T_n^{a*}=n(h_1\cdots h_q)^{1/2}I_n^{a*}/\sqrt{\hat{\sigma}_a^{*2}}$,其中 I_n^{a*} 和 $\hat{\sigma}_a^{*2}$ 与 I_n^a 和 $\hat{\sigma}_a^2$ 是相同的,除了用 $\hat{u}_i^*$ 来替代 $\hat{u}_i$。

(iv) 重复步骤(i)到(iii)较多次,比如 $B=399$ 次,然后构建 B 自举检验统计量 $\{T_{nj}^{a*}\}_{j=1}^B$ 的经验分布。这个自举经验分布用来近似检验统计量 T_n^a 的原分布。如果 $T_n^a>T_{n(\alpha B)}^{a*}$,则拒绝 H_0^a,其中 $T_{n(\alpha B)}^{a*}$ 是 $\{T_{nj}^{a*}\}_{j=1}^B$ 的上 α 分位点。

直观地,以上步骤(ii)和(iii)确保了条件于随机样本 $\{X_i,Y_i\}_{i=1}^n$,自举样本是通过原模型产生的,因为条件于 $\{X_i,Y_i\}_{i=1}^n$,u_i^* 具有零均值并且在步骤(iii)得到的自举统计量逼近检验统计量 T_n^a 的原分布,无论原假定成立与否。

358

接下来的定理证实了这种直觉并且证明了原始自举法的工作。

定理 12.3 在与定理 12.1 相同的条件下,除了我们没有施加原假定 H_0^a,则

$$\sup_{z\in\mathbb{R}}|P(T_n^{a*}\leqslant z|\{X_i,Y_i\}_{i=1}^n)-\Phi(z))|=o_p(1) \tag{12.8}$$

其中 $\Phi(\cdot)$ 是一个标准正态随机变量的累积分布函数。

定理 12.3 表明自举统计量 T_n^{a*} 依概率依分布收敛于一个 $N(0,1)$ 随机变量。见 Li,Hsiao 和 Zinn(2003),在文中给出另一种关于"依概率依分布收敛"的公式。

注意 $\Phi(z)$,一个服从 $N(0,1)$ 的随机变量的累积分布函数,是一个连续函数。利用 Polyā 定理(见 Bhattacharya 和 Rao(1986)),我们知道为了证明式(12.8),对于任意的 $z\in\mathbb{R}$ 的固定值,足以证明

$$|P(T_n^{a*}\leqslant z|\{X_i,Y_i\}_{i=1}^n-\Phi(z)|=o_p(1) \tag{12.9}$$

式(12.9)的证明与定理 12.1 的证明相类似,除了其需要使用一个不同的中心极限定理,因为条件于数据,T_n^{a*} 的首项是一个依赖于 i 和 j 的带有核函数的 U 统计量(是 U 统计量核函数而不是非参数核函数 $k(\cdot)$)。在这种情形下,我们可以使用 de Jong(1987)对二次型的中心极限定理来证明定理 12.3。如果想知道关于这个议题的更详细的讨论,可以参见 Fan,Li 和 Min(2006)以及 Hsiao 等(2007)。

在 Li 和 Wang(1998)中所报告的模拟表明 T_n^{a*} 检验在实际中表现良好,并且在合适的样本容量中(例如,$n=100$),其经验水平与其名义水平没有多大差异。Li 和 Wang 进一步说明,在 H_0 下,$|T_n^{a*}-T_n^a|=O_p(n^{-1/2}(h_1\cdots h_q)^{1/2})$。所以自举方法相比渐近正态逼近的确提供了一个对原分布更好的近似。再考虑当 $q=1,h=O(n^{-1/5})$ 的情况,自举法误差率近似值为 $O_p(n^{-1/2}h^{1/2})=O_p(n^{-3/5})$,远远小于渐近正态率误差近似值 $O_p(h^{1/2})=O_p(n^{-1/10})$。Fan 和 Linton(2003)给出

359

了自举法统计量 T_n^{a*} 精确度的进一步分析。

我们现在简要讨论加权函数 $M(x)=1$ 的选择。在这种情况下检验统计量被修正如下：

$$I_{n,m}^a = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq 1}^{n}\hat{u}_i\hat{u}_jK_{h,ij}/\hat{f}_{-i}(X_i) \tag{12.10}$$

其中 $\hat{f}_{-i}(X_i)=(n-1)^{-1}\sum_{j\neq i}^{n}K_{h,ij}$ 是 $f(X_i)$ 的去一核估计量。定理 12.1 被修正为

定理 12.4　在与定理 12.1 中给出的类似的条件下，进一步假定 X 有一个有界的支撑集，并且 $f(\cdot)$ 是有下界的，为 X 的支撑集中的一个正常数。那么在在 H_0^a 下，

$$T_{n,m}^a \stackrel{\text{def}}{=} n(h_1\cdots h_q)^{1/2}I_{n,m}^a/\sqrt{\hat{\sigma}_{a,m}^2}\xrightarrow{d}N(0,1)$$

其中 $\hat{\sigma}_{a,m}^2=\dfrac{2h_1\cdots h_q}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\hat{u}_i^2\hat{u}_j^2K_{h,ij}^2/\hat{f}_{-i}(X_i)^2$ 是 $\sigma_{a,m}^2=2\kappa^qE[\sigma^4(X)/f(X)]$ 的一致估计量。

定理 12.4 要求 X 有有界的支撑集，并且 $f(\cdot)$ 是有下界的，否则渐近方差 $E[\sigma^2(X)/f(X)]$ 可能不是有限的。或者，我们可以采用一个固定的修剪集来修剪掉 X_i 的较大值，这样我们就只从一个有界集里使用数据。修改自举法过程使得 $T_{n,m=1}^a$ 检验变得简便。产生 u_i^*、Y_i^* 和 $\hat{u}_i^*$ 的步骤(i)和(ii)保持不变。在步骤(iii)和(iv)中，以与 $T_{n,m}^a$（不是 T_n^a）相同的方法计算自举法统计量 $T_{n,m}^{a*}$，除了以 $\hat{u}_i^*$ 取代 $\hat{u}_i$。

或者，我们可以基于一个残差平方和的差，即 Ullah(1985) 提出的 $\text{SSR}_P-\text{SSR}_N$ 来检验参数回归函数形式的正确性。其中 $\text{SSR}_N=n^{-1}\sum_{i=1}^{n}\tilde{u}_i^2$，$\tilde{u}_i=Y_i-\hat{g}_{-i}(X_i)$，$\hat{g}_{-i}(X_i)$ 是 $E(Y_i/X_i)$ 的去一核估计值，$\text{SSR}_P=n^{-1}\sum_{i=1}^{n}\hat{u}_i^2$，$\hat{u}_i=Y_i-m(X_i,\hat{\gamma})$。Dette(1999) 对单变量 X 的情形推导出了一个基于残差平方检验的渐近原分布，而 Fan 和 Huang(2001) 以及 Fan 和 Li(2002) 考虑了一般多元变量 X 的情形下基于残差平方的检验。Fan 和 Li(2002) 证明，在 H_0 下，

$$\hat{T}_{n,\text{SSR}} \stackrel{\text{def}}{=} n(h_1\cdots h_q)^{1/2}[\text{SSR}_P-\text{SSR}_N+\hat{c}_n]/\hat{\sigma}_{\text{SSR}}\xrightarrow{d}N(0,1) \tag{12.11}$$

其中，当 $q\leqslant 3$ 时，$\hat{c}_n=\hat{c}_{2n}$；当 $q\geqslant 4$ 时，$\hat{c}_n=\hat{c}_{1n}+\hat{c}_{2n}$，　360

$$\hat{c}_{1n}=\frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}^{n}[m(x_i,\hat{\gamma})-m(x_j,\hat{\gamma})]^2K_{h,ij}^2/\hat{f}_{-i}(X_i)^2$$

$$\hat{c}_{2n}=\frac{1}{n^3}\sum_{i=1}\sum_{j\neq i}^{n}\hat{u}_i^2K_{h,ij}^2/\hat{f}_{-i}(X_i)^2$$

$$\hat{\sigma}^2_{\mathrm{SSR}} = \frac{2(h_1 \cdots h_q)}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{u}_i^2 \hat{u}_j^2 (\bar{K}_{h,ij} - 2K_{h,ij})^2 / \hat{f}_{-i}(X_i)^2$$

$$\hat{f}_{-i}(X_i) = \frac{1}{n-1} \sum_{j \neq i}^{n} K_{h,ij}$$

其中，$K_{h,ij} = \prod_{s=1}^{q} h_j^{-1} k((X_{js} - X_{is})/h_j)$，$\bar{K}_{h,ij} = \prod_{s=1}^{q} h_j^{-1} \bar{k}((X_{js} - X_{is})/h_j)$，$\bar{k}(v) = \int k(u+v)k(u)\mathrm{d}u$。

注意当 $q=1$ 时，式(12.11)退化为 Dette(1999)的定理 2.3 中给出的结果，除了式(12.11)只有一个中心项 $\hat{c}_{2n}$，因为式(12.11)使用了去一核估计量，与 Dette 没有采用去一估计量相比，它移走了一个中心项。

Fan 和 Li(2002)推荐使用自举法来更好地近似得到 $\hat{T}_{n,\mathrm{SSR}}$ 检验的零分布。令 $\hat{u}_i^* = Y_i^* - m(X_i, \hat{\gamma}^*)$ 为上面讨论的自举法残差，定义 $\tilde{u}_i^* = Y_i^* - \hat{Y}_i^*$，其中 $\hat{Y}_i^* = \sum_{j \neq i} Y_j^* K_{h,ij} / \sum_{j \neq i} K_{h,ij}$。那么自举法统计量为 $T^*_{n,\mathrm{SSR}} = n(h_1 \cdots h_q)^{1/2}[\mathrm{SSR}_P^* - \mathrm{SSR}_N^* + \hat{c}_n^*]/\hat{\sigma}^*_{\mathrm{SSR}}$，其中 SSR_P^*、SSR_N^*、$\hat{c}_n^*$ 和 $\hat{\sigma}^*_{\mathrm{SSR}}$ 是在 SSR_P、SSR_N、$\hat{c}_n$ 和 $\hat{\sigma}_{\mathrm{SSR}}$ 中分别以 $\tilde{u}_i^*$ 和 $\hat{u}_i^*$ 取代 $\tilde{u}_i$ 和 $\hat{u}_i$ 得到。

关于自举法理论的创造性工作见 Efron(1982)和 Hall(1992)；关于合适的自举重复次数的相关工作可参见 Davidson 和 MacKinnon(2000)，Andrews 和 Buchinsky(2002)。

12.1.2 混合数据

在这一节我们考虑 $X = (X^c, X^d)$ 的情况，其中 X^c 是一个 q 维连续变量，X^d 是一个 r 维离散变量。如果样本大小远远大于离散元的数量，而维度 q 很小，则可以使用第 3 章介绍的频率法来处理离散变量的存在。在这种情形下，检验统计量保持不变，除了核函数 $K_{h,ij}$ 需要以 $K_h(X_i^c, X_j^c)\mathbf{1}(X_i^d = X_j^d)$ 来替代，其中 $\mathbf{1}(X_i^d = X_j^d) = \prod_{s=1}^{r} \mathbf{1}(X_{is}^d = X_{js}^d)$。当 $X^d = X_j^d$ 时，该式等于 1，否则等于 0。经过这些修改，定理 12.1 和定理 12.2 的结果仍旧有效。

361

我们也可以如第 4 章所介绍的那样对离散变量进行平滑。那么检验统计量 I_n^a 需要以 $K_{\gamma,ij} = K_h(X_i^c, X_j^c) L(X_i^d, X_j^d, \lambda)$ 替代 $K_{h,ij}$ 来加以修正，其中 $L(X_i^d, X_j^d, \lambda) = \prod_{s=1}^{r} l(X_{is}^d, X_{js}^d, \lambda_s)$ 在第 4 章中已定义，$\gamma = (h, \lambda)$。可以通过(局部常数)最小二乘交错鉴定方法或者特定的方法来选择平滑参数$(h_1, \cdots, h_q, \lambda_1, \cdots, \lambda_r)$。令 $I_n^a(\hat{h}, \hat{\lambda})$ 表示得到的检验统计量，则定理 12.1 和定理 12.2 仍旧有效。例如，在

H_0 的条件下，定理 12.1 变成

$$T_n^a = n(\hat{h}_1 \cdots \hat{h}_q)^{1/2} I_n^a(\hat{h}, \hat{\lambda}) / \hat{\sigma}_a \xrightarrow{d} N(0,1)$$

其中

$$I_n^a \stackrel{\text{def}}{=} \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{u}_i \hat{u}_j K_{\hat{h}}(X_i^c, X_j^c) L(X_i^d, X_j^d, \hat{\lambda})$$

且

$$\hat{\sigma}_a^2 = \frac{2\hat{h}_1 \cdots \hat{h}_q}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \hat{u}_i^2 \hat{u}_j^2 K_{\hat{h}}(X_i^2, X_j^c)^2 L(X_i^d, X_j^d, \hat{\lambda})$$

Hsiao 等(2007)报告的模拟表明“原始自举法”得到了一个正确水平的检验。此外，交错鉴定方法对于缓慢变化的函数(低频率替换)倾向于选择较大的 h_s 值，而对迅速变化的函数(高频率替换)则倾向于选择相对较小的值。因此，基于交错鉴定方法的 T_n^a 检验比基于特定的窗宽选择规则的 T_n^a 检验，如 $h_s = X_{s,sd} n^{-1/(q+4)}$ 和 $\lambda_s = 0$ 具有更好的检验功效。

除了这一节讨论的检验之外，在文献中已经提出了很多可供选择的非参数检验用以检验参数回归函数形式的正确性。例如，可以参见 Azzalini，Bowman 和 Härdle(1989)，Wooldridge(1992)，Yatchew(1992)，Eubank 和 Spiegelman(1990)，Härdle 和 Mammen(1993)，Eubank 和 Hart(1992)，Ait-Sahalia，Bickel 和 Stoker(2001)，Ellison 和 Ellison(2000)，Liu，Stengos 和 Li(2000)，以及 Stengos 和 Sun(2001)。Tripathi 和 Kitamura(2003)，Chen，Härdle 和 Li(2003)，以及 Chen 和 Gao(2007)考虑了经验似然拟合优度检验。Hart(1997) 对基于非参数估计技术来检验参数回归函数形式的正确性提供了一个一般的处理。Horowitz(2006)构造了一个检验参数回归函数形式正确性的一致检验，针对在一个工具变量回归模型框架中的一般的非参数备择模型。我们将在第 17 章讨论非参数工具变量的估计。

362

12.2　概率密度函数等价的检验

在这一节我们考虑两个随机样本是否从同一分布抽取的检验问题。假定我们有两个独立同分布数据集 $\{X_i\}_{i=1}^{n_1}$ 和 $\{Y_i\}_{i=1}^{n_2}$，每一个都取自 q 维连续随机向量。假定 X 有概率密度函数 $f(\cdot)$，Y 有概率密度函数 $g(\cdot)$，我们要检验的原假定为

$$H_0^b: f(x) = g(x)，几乎所有\ x$$

我们构造一个检验统计量，基于 $f(\cdot)$ 与 $g(\cdot)$ 间的积分平方差，表达式由下式给出

$$
\begin{aligned}
I^b &= \int [f(x) - g(x)]^2 \mathrm{d}x \\
&= \int [f^2(x) + g^2(x) - 2f(x)g(x)]\mathrm{d}x \\
&= \int f(x)\mathrm{d}F(x) + \int g(x)\mathrm{d}G(x) - 2\int f(x)\mathrm{d}G(x)
\end{aligned}
$$

其中 $F(\cdot)$ 和 $G(\cdot)$ 分别是 X 和 Y 的累积分布函数。一个可行检验统计量可以通过以 $\hat{f}(\cdot)$、$\hat{g}(\cdot)$、$F_n(\cdot)$ 和 $G_n(\cdot)$ 分别取代 $f(\cdot)$、$g(\cdot)$、$F(\cdot)$ 和 $G(\cdot)$ 得到，其中 $\hat{f}(x) = n_1^{-1}\sum_{i=1}^{n_1} K_h(X_i, x)$，$\hat{g}(x) = n_2^{-1}\sum_{i=1}^{n_2} K_h(Y_i, x)$，$F_n(\cdot)$ 和 $G_n(\cdot)$ 分别为 $\{X_i\}_{i=1}^{n_1}$ 和 $\{Y_i\}_{i=1}^{n_2}$ 的经验累积分布函数。注意采用 Dirac 的 δ 函数方法（参见附录 A.40 的定义），对于任意（可测度的）函数 $M(\cdot)$，有 $\int M(x)\mathrm{d}F_n(x) = n_1^{-1}\sum_{j=1}^{n_1} M(X_j)$ 和 $\int M(x)\mathrm{d}G_n(x) = n_2^{-1}\sum_{j=1}^{n_2} M(Y_j)$。因此，我

363 们得到一个可行估计量：

$$
\begin{aligned}
I_b^n &= \int \hat{f}(x)\mathrm{d}F_n(x) + \int \hat{g}(x)\mathrm{d}G_n(x) - 2\int \hat{f}(x)\mathrm{d}G_n(x) \\
&= \left\{\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1} K_{h,ij}^x + \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2} K_{h,ij}^y - \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} K_{h,ij}^{xy}\right\}
\end{aligned}
$$

其中

$$
\begin{aligned}
K_{h,ij}^x &= \prod_{s=1}^{q} h_s^{-1} k((X_{is} - X_{js})/h_s) \\
K_{h,ij}^y &= \prod_{s=1}^{q} h_s^{-1} k((Y_{is} - Y_{js})/h_s) \\
K_{h,ij}^{x,y} &= \prod_{s=1}^{q} h_s^{-1} k((X_{is} - Y_{js})/h_s)
\end{aligned}
$$

为了推导 I_n^b 的渐近分布，需要下面的条件。

条件 12.2

(i) $f(\cdot)$ 和 $g(\cdot)$ 都是有界连续函数。

(ii) $k(\cdot)$ 是有界、非负二阶核函数。

定理 12.5 当条件 12.2 成立，在 H_0^b 下，我们有

$$
T_n^b \stackrel{\text{def}}{=} (n_1 n_2 h_1 \cdots h_q)^{1/2} \frac{(I_n^b - c_{n,b})}{\hat{\sigma}_b} \xrightarrow{d} N(0,1) \tag{12.12}
$$

其中

$$
c_{n,b} = \frac{k(0)^q}{h_1 \cdots h_q}\left[\frac{1}{n_1} + \frac{1}{n_2}\right]
$$

$$\hat{\sigma}_b^2 = \frac{h_1 \cdots h_q}{n_1 n_2}\left\{ \sum_{i=1}^{n_1}\sum_{j=1}^{n_1} \frac{(K_{h,ij}^x)^2}{(n_1/n_2)} + \sum_{i=1}^{n_2}\sum_{j=1}^{n_2} \frac{(K_{h,ij}^y)^2}{(n_2/n_1)} + 2\sum_{i=1}^{n_1}\sum_{j=1}^{n_2} (K_{h,ij}^{x,y})^2 \right\}$$

定理12.5的证明在12.6.3节中给出。注意在定理12.5中给出的$\hat{\sigma}_b^2$与Li(1996)给出的不同。习题12.4证明$\hat{\sigma}_b^2$渐近等价于Li给出的方差估计值。

可以证明T_n^b确实是一个一致性检验,当H_0^b不成立时,它在各方向都具有检验功效。也就是说,如果H_0不成立,则

$$P(T_n^b > C) \to 1, \text{任意正常数} C \tag{12.13}$$

T_n^b是单侧检验,因此在实践中如果$T_n^b > Z_\alpha$,则在置信水平α上拒绝H_0^b,Z_α是标准正态变量的上α分位点。

364

注意到I_n^b检验有一个非零中心项$c_{n,b}$,为了得到一个渐近零均值检验(在H_0^b下)必须将它减掉。在定理12.5的证明中我们知道中心项来自I_n^b中$i=j$项的出现。Li(1996)也考虑了一个中心值不受限制的检验统计量,这个检验统计量是通过移除I_n^b中的$i=j$项得到。定义

$$I_{1n}^b = \left\{ \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j\neq i,j=1}^{n_1} K_{h,ij}^x + \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j\neq i,j=1}^{n_2} K_{h,ij}^y - \frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j\neq i,j=1}^{n_2} K_{h,ij}^{x,y} \right\} \tag{12.14}$$

通过对定理12.5的证明,我们立即得到下述推论:

推论12.1 在与定理12.5给出的相同条件下,我们有

$$H_0^b: T_{1n}^b \overset{\text{def}}{=} (n_1 n_2 h_1 \cdots h_q)^{1/2} I_{1n}^b / \hat{\sigma}_b \xrightarrow{d} N(0,1) \tag{12.15}$$

Li(1996)报告的模拟证明了中心值不受限制的T_{1n}^b检验比T_n^b检验具有稍微好一点的有限样本性质和功效;但是,中心值不受限制的T_{1n}^b检验存在一个问题。在有限样本里,这个统计量的值依赖于X_i和$Y_j(i=1,\cdots,n_1;j=1,\cdots,n_2)$的排列次序。要看到这一点需要关注

$$I_{1n}^b = I_n^b - \left(\frac{1}{n_1} + \frac{1}{n_2}\right)\frac{k(0)^q}{h_1 \cdots h_q} - \frac{2}{n_1 n_2}\sum_{i=1}^{\min\{n_1,n_2\}} K_{ii}^{x,y} \tag{12.16}$$

尽管I_n^b不随着数据的排列次序而改变,但式(12.16)的第三项

$$2\,(n_1 n_2)^{-1} \sum_{i=1}^{\min\{n_1,n_2\}} K_{ii}^{x,y}$$

依赖于数据是如何排序的。这意味着I_{1n}^b必定依赖于数据的排列次序。式(12.16)的最后一项是渐近可以被忽略的,因此不同的排列次序并不影响I_{1n}^b的渐近分布,但在有限样本应用中可能会影响I_{1n}^b检验过程的结果。

因为横截面数据集没有自然排序,而且随机地对数据进行重新排序可能得到不同的检验结果,在实践中当使用T_{1n}^b检验时需要小心。相反,T_n^b检验并不随着数据排列次序的不同而改变。

365　式(12.12)中的检验与 Mammen(1996)提出的检验基本上是相同的，除了 Mammen 的检验要求计算核函数的双重卷积和四重卷积。Anderson，Hall 和 Titterington(1994)也提出了一个类似 T_n^b 的检验，但有一个固定平滑参数(非平滑检验)，这使得他们的检验统计量为一个非标准分布。他们建议使用自举法来近似他们检验统计量的零分布。我们将在第 13 章讨论非平滑检验。

Mammen(1992)建议使用自举法来更好地近似检验统计量的零分布。这项工作由从混合数据中有放回随机抽样实现。令 Z_i 表示对混合数据的第 i 个样本实现，从 $\{Z_i\}_{i=1}^{n_1+n_2}$ 中有放回随机抽取 n_1 个观测值，称该样本为 $\{X_i^*\}_{i=1}^{n_1}$。然后从 $\{Z_i\}_{i=1}^{n_1+n_2}$ 有放回随机抽取 n_2 个观测值，称其为 $\{Y_i^*\}_{i=1}^{n_2}$。除了分别用 X_i^* 和 Y_i^* 取代 X_i 和 Y_i 外，用与 T_n^b 相同的方法计算检验统计量 T_n^{b*}。这个步骤需要重复很多次，比如 $B=399$ 次，得到的自举统计量 $\{T_{nj}^{b*}\}_{j=1}^{B}$ 称为检验统计量的自举样本。如果 $T_n^b > T_{n(\alpha B)}^{b*}$，则拒绝原假定 H_0^b，这里 $T_{n(\alpha B)}^{b*}$ 是 $\{T_{nj}^{b*}\}_{j=1}^{B}$ 的上 α 分位点。比如，在 $\alpha=5\%$ 的显著性水平和 $B=399$ 次时，将 399 个自举统计量从小到大排列，$T_{n(1)}^{b*}\leqslant T_{n(2)}^{b*}\leqslant\cdots\leqslant T_{n(399)}^{b*}$，则 $T_{n(1-(\alpha(B+1))}^{b*}=T_{n(380)}^{b*}$。或者可以计算由 m/B 给出的自举 P 值，这里 m 是自举统计量(T_n^{b*})超过原始统计量 T_n^b 的数目。

对于两个概率密度函数是否相同的其他相关检验包括 Fan 和 Gencay(1993)。对于检验单变量密度是否相等的相关测度熵检验参见 Robinson(1991)，Maasoumi 和 Racine(2002)，以及 Granger，Maasouni 和 Racine(2004)。

12.3　关于回归函数更多的检验

12.3.1　参数回归模型的 Härdle 和 Mammen 检验

366　Härdle 和 Mammen(1993)提出了一个参数回归函数形式的一致性检验，基于模拟的样本 I_{HM}：

$$I_{\mathrm{HM}}=\int[\hat{g}(x)-m(x,\hat{\gamma})]^2 w(x)\,\mathrm{d}x$$

其中 $w(x)$ 是一个加权函数。

如果选择 $w(x)=f^2(x)$，且用 $\hat{f}(x)=n^{-1}\sum_{i=1}^{n}K_h(X_j,x)$ 代替 $f(x)$，则以上统计量变为

$$I_{n,\mathrm{HM}}=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}[Y_i-m(X_i,\hat{\gamma})][Y_i-m(X_j,\hat{\gamma})]\times\int K_h(X_i,x)K_h(X_j,x)\,\mathrm{d}x$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\hat{u}_i\hat{u}_j\bar{K}_h(X_i,X_j) \tag{12.17}$$

其中 $\hat{u}_i = Y_i - m(X_i,\hat{\gamma})$，$\bar{K}_h(X_i,X_j) = \prod_{s=1}^{q} h_s^{-1}\bar{k}(X_{is} - X_{js})/h_s)$，且 $\bar{k}(v) = \int k(u)k(v+v)\mathrm{d}u$ 是来自 k 的双重卷积核。

对比 I_n^a 和 $I_{n,\mathrm{HM}}$，我们发现二者唯一的区别在于后者包括双重求和中的 $i=j$ 项，并且 K 由 $\bar{K}$ 代替。因此，由定理 12.1 的结果，我们可以立即得出以下推论。

定理 12.6 在与定理 12.1 相同的条件下，在 H_0^a 下，

$$n(h_1\cdots h_q)^{1/2}(I_{n,\mathrm{HM}} - c_{n,a})/\tilde{\sigma}_a \xrightarrow{d} N(0,1)$$

其中 $c_{n,a} = \frac{\bar{k}(0)^q}{nh_1\cdots h_q}\sum_{i=1}^{n}\hat{u}_i^2$ 且 $\tilde{\sigma}_a^2 = \frac{2h_1\cdots h_q}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}^{n}\hat{u}_i^2\hat{u}_j^2\bar{K}_h(X_i,X_j)^2$

证明：注意 $I_{n,\mathrm{HM}} - c_{n,a}$ 等于 I_n^a，除了在 I_n^a 中的核 K 被卷积核 $\bar{K}$ 替代。$\bar{K}$ 也是一个二阶核。因此，定理 12.6 由定理 12.1 得出。 □

如果在 $I_{n,\mathrm{HM}}$ 中选择 $w(x) = f(x)$，并且用样本均值替代总体均值，将给出 Härdle 和 Mammen(1993)检验的另一形式，

367

$$\tilde{I}_{n,\mathrm{HM}} = n^{-1}\sum_{i=1}^{n}[\hat{g}(X_i) - m_h(X_i,\hat{\gamma})]^2 \tag{12.18}$$

关于非参数检验的早期文献依赖经验法则的窗宽选择方法来建造一致模型设定检验(参见 Härdle 和 Mammen(1993)，Zheng(1996)，Li 和 Ming(1998))。遗憾的是，这种选择平滑参数的专门方法被证实会导致功效损失，由此带来更多近期发展，参见 Hsiao 等(2007)，他建议用数据驱动的交错鉴定方法来挑选平滑参数，Horowitz 和 Spokoiny(2001)提出基于式(12.18)的自适应最优速率检验，这正是下一节的主题。

12.3.2 自适应最优速率检验

平滑参数选择是非参数核分析的一个关键组成部分，在构造非参数模型设定检验时它也极其重要。对于非参数核检验，实证大小和功效都对窗宽值很敏感。当使用自举法时，非参数检验的实证大小通常对平滑参数中一些变化并不十分敏感。然而，实证功效对这些变化是敏感的。因此，一个关键的问题为如何选择最佳的平滑参数使得最终的检验对于原模型中所有偏离都有强大的功效。直觉上，人们可以认同对于低频率的数据应该用相对较大的窗宽来精确估计对原模型的偏离。然而，对于高频率数据，大的窗宽值会过度平滑数据并使原模型中的离差变得模糊。因此，对高频率数据必须选择一个相对较小的窗宽值(更多讨论详见 Fan 和 Li(2000))。

Horowitz 和 Spokoiny(2001)提出了一个具有对低频和高频数据都拥有高功效优势的自适应检验统计量。他们考虑了一个在式(12.18)中给出的基于一个学生化(studentized)版本 $I_{n,\mathrm{HM}}$ 的检验。在 X_i 为非随机的假定下,很容易得到:

368

$$N_h \overset{\text{def}}{=} E(\tilde{I}_{n,\mathrm{HM}}) = \sum_{i=1}^{n} a_{ii,h}\sigma^2(X_i)$$

$$V_h^2 \overset{\text{def}}{=} \mathrm{var}(\tilde{I}_{n,\mathrm{HM}}) = 2\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij,h}^2\sigma^2(X_i)\sigma^2(X_j)$$

其中 $a_{ij,h} = \sum_{l=1}^{n} w_{il}w_{lj}$, $w_{ij} \equiv K((X_i - X_j)/h)/\sum_{l=1}^{n} K((X_i - X_l)/h)$。因此,一个学生化检验由下式给出:

$$\tilde{T}_h = \frac{\tilde{I}_{n,\mathrm{HM}} - N_h}{V_h}$$

且一个可行性检验为

$$\hat{T}_h = \frac{\tilde{I}_{n,\mathrm{HM}} - \hat{N}_h}{V_h}$$

其中 $\hat{N}_h = \sum_{i=1}^{n} a_{ii,h}\hat{\sigma}^2(X_i)$,且

$$\hat{V}_h^2 = 2\sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij,h}^2\hat{\sigma}^2(X_i)\hat{\sigma}^2(X_j)$$

$\hat{\sigma}^2(x)$是在假定 H_0^a 或 H_1^a 下的 $\sigma^2(x)$一致估计量(参见 Horowitz 和 Spokoiny 构造 $\hat{\sigma}^2(x)$的详细讨论)。

为了得到自适应检验,Horowitz 和 Spokoiny(2001)提出了使用几个不同的平滑参数值 $h \in H_n = \{h_{n1}, h_{n2}, \cdots, h_{nm}\}$ 的方法,m 是有限数值。H_n 的一个例子为 $H_n = \{h_s = h_{s,\max}a^k : h_s \geq h_{s,\min}, s = 1, \cdots, q; k = 0, 1, 2, \cdots\}$,其中 $0 < h_{s,\min} < h_{s,\max}$ 且 $0 < \alpha < 1$。继而检验统计量选取所有 h 值中的最大值,即

$$\hat{T} = \max_{\{h \in H_n\}} T_h \tag{12.19}$$

Horowitz 和 Spokoiny(2001)建议使用自举法来近似得到检验统计量 $\hat{T}$的临界值。自举法步骤为:

(i) 生成 $Y_i^* = m(X_i, \hat{\gamma}) + u_i^*, i = 1, \cdots, n$,其中 u_i^* 由 $N(0, \hat{\sigma}^2(X_i))$随机抽取。

369

(ii) 运用自举样本$\{X_i, Y_i^*\}_{i=1}^n$估计 γ 和 $\sigma^2(X_i)$,得到的相应估计量记为 $\hat{\gamma}^*$ 和 $\hat{\sigma}^{2*}(X_i)$。用在式(12.19)中定义 $\hat{T}$ 的相同方法计算 $\hat{T}^*$,但以 Y_i^*、$\hat{\gamma}^*$ 和 $\hat{\sigma}^{2*}(X_i)$来替代 Y_i、$\hat{\gamma}$ 和 $\hat{\sigma}^2(X_i)$。

(iii) 令 t_α 表示显著性水平为 α 检验的渐近零临界值。用重复步骤(i)和

(ii)多次得到的 $\hat{T}^*$ 经验分布的 $1-\alpha$ 分位数估计 t_α。

Horowita 和 Spokoiny(2001)证明了以下结论：

定理 12.7　在 Horowita 和 Spokoiny(2001)给出的正则性条件下，令假定 H_0^e 为真，则

$$\lim_{n\to\infty} P\left(\hat{T}^* > t_\alpha\right) = \alpha$$

Horowita 和 Spokoiny(2001)中的模拟证明他们的检验比 T_n^a检验具有更好的功效，尤其在高频率数据方面。

12.3.3　参数单指数模型检验

第 8 章中介绍的参数单指数模型具有以下形式：

$$Y_i = g\left(X'_i\beta\right) + u_i, E\left(u_i \mid X_i\right) = 0$$

其中 $g(\cdot)$是已知函数，β 是一个未知参数向量。例如，如果 $g(\cdot) = \Phi(\cdot)$，$\Phi(\cdot)$是一个标准正态变量的累积分布函数，我们得到流行的 Probit 模型。

因此原假定是 H_0^c:$E\left(Y_i \mid X_i\right) = g\left(X_i'\beta\right)$，或等价地 $E\left(u_i \mid X_i\right) = 0$，其中 $u_i = Y_i - g\left(X_i'\beta\right)$。按照 12.1.1 节的分析，我们可以构建一个基于 $E[uE(u \mid X'\beta)f(X'\beta)]$的检验统计量：

$$I_n^c = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i, j=1}^{n} \hat{u}_i \hat{u}_j K_h(X_i, X_j) \tag{12.20}$$

其中 $\hat{u}_i = Y_i - g\left(X_i'\hat{\beta}\right)$，$\hat{\beta}$ 是基于参数原模型的 β_0 的$\sqrt{n}$一致估计量。它可以是一个非线性最小二乘估计量，或者在 Probit 或 Logit 模型情况下，它可以是 β_0 的最大似然估计量。

条件 12.3　370

(i) 如果 $g(X)$ 和 u 都有有限的四阶矩。$g(\cdot)$为二次可微，$\sigma^2(v) = E(u^2 \mid x\beta = v)$连续。

(ii) $k(\cdot)$是非负的二阶核函数，并 m 次可微($m\geq 3$ 是一个整数)。

(iii) 当 $n\to\infty$ 时，$h\to 0$，$n^{(m-2)}h^{(2m-1)}\to\infty$。

(iv) 在 H_0^c 下，$\hat{\beta}-\beta = O_p\left(n^{-1/2}\right)$。

Fan 和 Liu(1997)证明了下面的结论。

定理 12.8　在条件 12.3 和假定 H_0^c 下，可得到

$$nh^{1/2} I_n^c / \hat{\sigma}_c^2 \xrightarrow{d} N(0,1)$$

其中 $\hat{\sigma}_c^2 = \dfrac{2h}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i}^{n} \hat{u}_i^2 \hat{u}_j^2 K_h\left(\left(X_i - X_j\right)'\hat{\beta}\right)^2$。

如同 12.1.1 节的情形，我们也可以用自举法更好地逼近 I_n^c 的零分布。

使用类似 Härdle 和 Mammen(1993)的方法,Horowitz 和 Härdle(1994)提出一个方法检验相对半参数的一般参数指数模型,该方法也能用作 H_0^c 的一致性检验。Hristache 等(2001)提出一个可供选择的针对单指数模型的拟合优度检验。

12.3.4 非参数遗漏变量检验

"显著性检验"可能是应用回归分析中使用最频繁的检验,而且经常被用作证实或推翻经济理论。稳健的参数推断依赖于潜在数据生成过程的正确函数设定。然而,参数框架中错误设定的可能性不能被忽视,尤其是考虑到应用研究人员倾向于在简洁和易处理的基础上选择参数模型这个事实。给定非参数方法在比参数方法需要的约束假定少很多的情况下是一致的,非参数核框架中的显著性检验因此具有明显的吸引力。

371 在这一节,我们为应用研究者概述一种在一个完全非参数和稳健框架中的检验假定的方法,从而回应潜在的批评,即一个特定的结果是由于任意的参数设定带来的。

令 $X \in \mathbb{R}^q$ 为连续随机变量的一个 $q \times 1$ 维向量,划分 $X = (W, Z) \in \mathbb{R}^p \times \mathbb{R}^{q-p}(1 \leq p < q)$。原假定为 Y 的条件均值不取决于 Z,即

$$H_0^d: E(Y \mid w, z) = E(Y \mid w) \quad \text{a.e.}$$

定义 $u = Y - E(Y|W)$,则在 H_0^d 下有几乎处处 $E(u|X) = 0$,我们可以构建一个基于 $E\{uf_w(W)E[uf_w(W)|X]f(X)\}$ 的检验统计量,其中 $f_w(\cdot)$ 和 $f(\cdot)$ 分别为 W 和 X 的概率密度函数,令 $\hat{f}_{w_i}$ 和 $\hat{Y}_i$ 分别表示 $f_w(W_i)$ 和 $E(Y_i|W_i)$ 的去一核估计量,即

$$\hat{f}_{w_i} = \frac{1}{(n-1)} \sum_{j \neq i}^{n} K_{h_w}(W_i, W_j)$$

$$\hat{Y}_i = \frac{1}{(n-1)} \sum_{j \neq i}^{n} Y_j K_{h_w}(W_i, W_j) / \hat{f}_{w_i}$$

$$K_{h_w}(W_i, W_j) = \prod_{s=1}^{p} h_{w,s}^{-1} k_w \left(\frac{W_{is} - W_{js}}{h_{w,s}} \right)$$

其中 $k_w(\cdot)$ 是单变量核函数,$h_{w,s}$ 是与 $W_{i,s}$ 相联系的平滑参数。一个可行的检验统计量可以给出为:

$$I_n^d = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} (Y_i - \hat{Y}_i)\hat{f}_{w_i}(Y_j - \hat{Y}_j)\hat{f}_{w_j} K_h(X_i, X_j)$$

其中 $K_h(X_i, X_j) = \prod_{s=1}^{q} h_s^{-1} k((X_{is} - X_{js})/h_s)$ 是用来估计 $E(uf_w(W)|X)f(X)$ 的核函数。

以下条件用于推导出 I_n^d 的渐近分布。

372

条件 12.4

(i) 定义 $u_i = Y_i - E(Y_i | W_i)$，令 $\mu_4(x) = E(u_i^4 | x)$，f_z，$f(\cdot)$，σ^2 和 μ_4 均满足 Lipschitz 条件，即 $|m(u+v) - m(u)| \leqslant C|v|$。令 $E(Y|x)$ 为 ν 次可微，$E(Y|w)$ 为一个连续函数。

(ii) 令 $k_w(\cdot)$ 为一个 ν 阶核，令 $k(\cdot)$ 为一个非负二阶核。

(iii) 当 $n \to \infty$ 时，$h_{w,s} \to 0 (s=1,\cdots,p)$，$h_s \to 0 (s=1,\cdots,q)$，$nh_{w,1}\cdots h_{w,p} \to \infty$，$nh_1\cdots h_q \to \infty$，$n(h_1\cdots h_q)^{1/2}\left(\sum_{s=1}^{p} h_{w,s}^{2\nu}\right) \to 0$ 且 $(h_1\cdots h_q)/(h_{w,1}^2\cdots h_{w,p}^2) \to 0$。

条件 12.4(i) 包含对未知函数的标准平滑性条件。条件 12.4(ii) 假定使用一个 ν 阶的核估计原模型。这意味着原模型的均方误差是 $O\left(\sum_{s=1}^{q} h_{w,s}^{2\nu} + (nh_{w,1}\cdots h_{w,q})^{-1}\right)$ 阶的。一致模型设定检验通常包含一个二阶退化 U 统计量，并且这样一个 U 统计量的标准化常数是 $(n(h_1\cdots h_q)^{1/2})^{-1}$。条件 12.4(iii) 假定来自原模型的均方误差的阶比 $(n(h_1\cdots h_q)^{1/2})^{-1}$ 小，即 $n(h_1\cdots h_q)^{1/2}\left(\sum_{s=1}^{p} h_{w,s}^{2\nu} + (nh_{w,1}\cdots h_{w,q})^{-1}\right) = o(1)$。注意如果 $p \leqslant q/2$，最后一个条件意味着 h_s 的阶比 $h_{w,s}$ 小，则意味着我们对关于原模型的备择回归模型必定平滑不足(undersmooth)。

定理 12.9 在条件 12.4 及 H_0^d 假定下，我们有

$$T_n^d \stackrel{\text{def}}{=} (nh_1\cdots h_q)^{1/2} I_n^d / \hat{\sigma}_d \xrightarrow{d} N(0,1)$$

其中 $\hat{\sigma}_d^2 = \dfrac{2h_1\cdots h_q}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{u}_i^2 \hat{f}_{w_i}^2 \tilde{u}_j^2 \hat{f}_{wj}^2 K_h(X_i, X_j)^2$，$\tilde{u}_i = Y_i - \hat{Y}_i$，而其中 $K_h(X_i, X_j) = \prod_{s=1}^{q} h_s^{-1} k((X_{is} - X_{js})/h_s)$。

可以证明如果 H_0^d 不成立，则 $n \to \infty$ 时，$P(T_n^d > C) \to 1$，因此 T_n^d 是一个一致性检验。

检验统计量 T_n^d 在原假定下是渐近标准正态分布。然而，Li(1999)以及 Lavergne 和 Vuong(2000)报告的模拟结果揭示出正态近似在小样本和中样本中效果并不好，并且检验统计量也依赖于 $h_{w,s}$ 和 h_s 的两个平滑参数集，并且对平滑参数的选择很敏感。因此 Lavergne 和 Vuong 建议用一个 T_n^d 检验的修正版本，该版本在有限样本上的效果比 I_n^d 检验更好。或者，也可以使用自举法来更好地近似 I_n^d 的零分布。

373

为了在实践中应用 T_n^d 检验，首先，回归元 X 必须分割成两部分，W 和 Z。我

们推荐使用局部常数最小交错鉴定法来帮助确定哪些变量属于 Z。令 $h_s = c_s X_{s,sd} n^{-1/(4+q)}$（$X_{s,sd}$是$\{X_{is}\}_{i=1}^n$的样本标准离差），则我们可以搜索 c_s 来最小化交错鉴定函数。从第 2 章定理 2.4 的结果我们知道当 x_s 是一个不相关回归元时，c_s 应该发散到（正）无穷。因此，当 c_s 很大时，比如 c_s 超过 n^{4+q}，x_s 可以选为 Z 的一部分（等同于 $h_s > X_{s,sd}$）。

T_n^d的自举法过程

我们使用 u_i^* 来表示原始自举误差，该误差来自拟合残差 $\tilde{u}_i = Y_i - \hat{Y}_i$，即以概率 $r = (1+\sqrt{5})/(2\sqrt{5})$ 有 $u_i^* = [(1-\sqrt{5})/2]\tilde{u}_i$；以概率 $1-r$ 有 $u_i^* = [(1+\sqrt{5})/2]\tilde{u}_i$。

自举法检验统计量通过以下步骤得到：

(i) 使用原始自举误差 u_i^* 得到 $Y_i^* = \hat{Y}_i + u_i^*$，然后通过下式得到 $E^*(Y_i^* \mid w_i) f_{w_i}$的核估计量：

$$\hat{Y}_i^* \hat{f}_{w_i} = (n-1)^{-1} \sum_{j \neq i}^{n} Y_j^* K_{h_w, ij}$$

估计的密度加权自举残差为 $\tilde{u}_i^* \hat{f}_{w_i} = (Y_i^* - \hat{Y}_i^*) \hat{f}_{w_i} = Y_i^* \hat{f}_{w_i} - \hat{Y}_i^* \hat{f}_{w_i}$。

(ii) 计算自举法检验统计量

$$I_n^{d*} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \tilde{u}_i^* \hat{f}_{w_i} \tilde{u}_j^* \hat{f}_{w_j} K_{h,ij}$$

则估计的渐近方差为

$$\hat{\sigma}_d^{*2} = 2(n^2 h_1 \cdots h_q)^{-1} \sum_{i=1}^{n} \sum_{j \neq i}^{n} \tilde{u}_i^{*2} \hat{f}_{w_i}^2 \tilde{u}_j^{*2} \hat{f}_{w_j}^2 K\left(\frac{X_i - X_j}{h}\right)^2$$

374 由此标准化的自举法统计量由 $T_n^{d*} = n(h_1 \cdots h_q)^{\frac{1}{2}} I_n^{d*} / \hat{\sigma}_d^*$ 给出。

(iii) 多次重复步骤(i)和步骤(ii)，如 $B = 399$ 次，得到 B 检验统计量 T_n^{d*} 的经验分布。令 $T_{n,\alpha}^{d*}$表示自举法分布的 α 分位点。如果 $T_n^d > T_{n,\alpha}^{d*}$，我们可以在 α 的显著性水平下拒绝原假定。

Gu 和 Li(2006) 报告的模拟结果显示上述自举法检验成功地克服了使用渐近正态近似时出现的置信水平失真。估计的置信水平接近于名义值，结果对平滑参数的选择相对不敏感。

构造统计量 I_n^d 时使用密度加权避免了随机分母，因而简化了渐近分析。类似在 12.1.1 节中讨论的统计量 $I_{n,m}^a$，也可以使用一个非密度加权形式，它将导致以下修正的检验统计量：

$$I_{n,m}^d = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} (Y_i - \hat{Y}_i)(Y_j - \hat{Y}_j) K_h(X_i, X_j) / \hat{f}(X_i) \quad (12.21)$$

则定理12.9被修改为如下：

定理12.10 在定理12.9的相似条件下，同样假定$f_e(\cdot)$和$f(\cdot)$在W和X的支撑上都是有下界的，且下界为正的常数。则在H_0^d假定下，我们有：

$$T_{n,m}^d \stackrel{\text{def}}{=} (nh_1\cdots h_q)^{1/2} I_{n,m}^d / \sqrt{\hat{\sigma}_{d,m}^2} \xrightarrow{d} N(0,1)$$

其中$\hat{\sigma}_{d,m}^2 = (2h_1\cdots h_q/n^2)\sum_{i=1}^{n}\sum_{j=1}^{n} \tilde{u}_i^2 \tilde{u}_j^2 K_h(X_i,X_j)^2/\hat{f}(X_i)^2$。

自举法统计量$T_{n,m}^{d*}$可以通过用$\tilde{u}_i^*$替换$\tilde{u}_i$从$T_{n,m}^d$得到，其中$\tilde{u}_i^* = Y_i^* - \hat{Y}_i^*$，$Y_i^*$和构造（密度加权）统计量$T_{n,m}^{d*}$时讨论的一样，有$\hat{Y}_i^* = (n-1)^{-1}\sum_{j\neq i}^{n} Y_i^* K_{h_w,ij}/\hat{f}_{w_i}$。

相关工作包括Lavergne和Vuong(1996)，他们考虑了在非嵌套式回归模型框架中选择非参数回归元的问题，并且Donald(1997)提出一种在多元非参数关系下选择因子的非参数检验，其与本节讨论的非参数显著性检验很相似。

12.3.5 分类变量的显著性检验

375

在这一节中我们考虑在一个非参数回归模型中分类变量的显著性检验问题。上一节讨论的检验可以被扩展到这样的情形，使用传统的非参数频率估计方法的分类变量的显著性检验。但是，这个检验在有限样本下可能会丧失检验功效，因为传统的频率方法把样本划分为许多“离散单元”或者子样本，而只使用在每个单元里的观测值来产生一个非参数估计。这样的效率损失是很令人遗憾的，因为在原假定下，某些离散变量是不相关的回归元，因此需要从回归模型中移除，也就是说，相应的离散单元需要被平滑掉，而不是把样本划分成不同的离散单元。

考虑一个有混合了分类和连续回归元的非参数回归模型，我们关注的是要检验是否某些分类回归元是不相关的，即多余的。将可能是多余的分类解释变量记为z。记X为回归模型中其余的解释变量，Y为被解释变量。则原假定可以写成：

$$H_0^c: E(Y\mid x,z) = E(Y\mid x)\ \text{a.e.}$$

备择假定是对假定H_0^c的否定，即在具有正测度集上有$H_1^c: E(Y|x,z)\neq E(Y|x)$。我们允许$X$同时包含分类（离散）变量和连续变量。首先，考虑$Z$是一个一元分类变量的情况。然后，在本节末讨论$Z$为多元的情况。

令$g(x)=E(Y|x)$，$m(x,z)=E(Y|x,z)$。原假定为几乎处处$m(x,z)=g(x)$。假定一元Z取c个不同值$\{0,1,2,\cdots,c-1\}$。如果$c=2$，则Z是一个0－1的虚拟变量，这在实践中可能是最经常遇到的情况。

注意我们已经假定 Z 是个一元分类变量,如果 Z 是一个有序(ordinal 或 ordered)分类变量,则当 $Z_i=Z_j$ 时,Z 的核函数为 $l(Z_i,Z_j,\lambda_Z)=1$;当 $Z_i\neq Z_j$ 时,$l(Z_i,Z_j,\lambda_Z)=\lambda_z^{|Z_i-Z_j|}$。如果 Z 是名义变量(即无序变量),则当 $Z_i=Z_j$ 时,$l(Z_i,Z_j,\lambda_Z)=1$,否则为 λ_Z。

采用同样的简化记号 $K_{h,ij}=\prod_{s=1}^{q}h_s^{-1}k((X_{is}^c-X_{js}^c)/h_s)$,$L_{\lambda,ij}=\prod_{s=1}^{r}l(X_{is}^d,X_{js}^d,\lambda_s)$,$l_{\lambda_z,ij}=\prod_{s=1}^{r}l(Z_i,Z_j,\lambda_z)$ 来表示核函数。根据第 4 章介绍的最小二乘交错鉴定方法选择$(h_1,\cdots,h_q,\lambda_1,\cdots,\lambda_r,\lambda_z)$。

376

使用 $\hat{h}_s$、$\hat{\lambda}_s$、$\hat{\lambda}_z$ 来分别表示交错鉴定选择的 h_s、λ_s 和 λ_z。当 H_1^e 成立时,根据定理 4.7 我们可知 $\hat{\lambda}_z=O_p(n^{-2/(q+4)})$,因为一致非参数估计要求当 $n\to\infty$ 时 $\hat{\lambda}_z\to 0$。然而,当 H_0^e 成立时,$\hat{\lambda}_z$ 将收敛到它的上界值 1。在这个情况下,交错鉴定方法将平滑掉不相关的回归元 Z,这将比传统频率方法更有效,它在即使离散变量不相关的情况下把样本划分为很多子集。

注意对于所有的 X 及 $l=1,\cdots,c-1$,原假定 H_0^e 等同于 $m(x,z=l)=m(x,z=0)$。检验统计量为 I^e 的一个估计值。

$$I^e=\sum_{l=1}^{c-1}E\{[m(x,z=l)-m(x,z=0)]^2\}$$

显然有 $I^e\geqslant 0$,当且仅当 H_0^e 成立时,$I^e=0$。因此,I^e 被看成是检验 H_0^e 成立与否的适当指标。一个可行的检验统计量由下式给出:

$$I_n^e=\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{c-1}[\hat{m}(X_i,Z_i=l)-\hat{m}(x,z=0)^2] \tag{12.22}$$

其中

$$\hat{m}(X_i,Z_i=l)=\frac{\sum_{j=1}^{n}Y_jW_{\hat{h},ij}L_{\hat{\lambda},ij}l_{Z_j,z=l,\hat{\lambda}_z}}{\sum_{j=1}^{2}W_{\hat{h},ij}L_{\hat{\lambda},ij}l_{Z_j,z=l,\hat{\lambda}_z}} \tag{12.23}$$

很容易证明,I_n^e 是 I^e 的一个一致估计量。因此,在 H_0^e 假定条件下,I_n^e 依概率收敛于 0;在 H_1^e 假定条件下,I_n^e 依概率收敛于 $I>0$。在实际中,如果 I_n^e 取一个过大的值,则拒绝假定 H_0^e。

为了处理多元 Z 的情形,推广检验统计量式(12.22)是直接的。假定 Z 为 d 维变量。令 z_s 和 z_{is} 分别表示 Z 和 Z_i 的第 s 个元素。假定 z_s 在 $\{0,1,\cdots,c_{z,s}-1\}(s=1,\cdots,d)$ 中取 $c_{z,s}$ 个不同的值。对于多元 Z,检验统计量 I_n^e 变成

$$I_n^e = \frac{1}{n}\sum_{i=1}^{n}\sum_{z}[\hat{m}(X_i,z) - \hat{m}(X_i,z_1 = 0,\cdots,z_r = 0)]^2 \qquad (12.24)$$

其中 $\sum_z$ 表示 $z \in \prod_{s=1}^{q}\{0,1,\cdots,c_{z,s}-1\}$ 所有可能值的加总。$\hat{m}(X_i,z)$的定义类似式(12.3),只是一元核 $\bar{l}(Z_i,Z_j,\hat{\lambda}_z)$ 由乘积核 $\prod_{s=1}^{d} l(Z_{s,i},Z_{s,j},\hat{\lambda}_{z,s})$ 取代,$\hat{\lambda}_{z,s}$ 定义为与 $z_s(s=1,\cdots,d)$ 的相联系的平滑参数 $\lambda_{z,s}$ 的交错鉴定值。我们现在把关注转向自举法,来逼近检验统计量的有限样本零分布。 377

I_n^e 检验的渐近分布相当复杂(它依赖于 $\hat{\lambda}_z$ 的零分布)。Racine 等(2006)建议使用自举法来逼近 I_n^e 的零分布。注意在这个检验问题中,我们不能从 $\{Y_i,X_i,Z_i\}_{i=1}^n$ 中再取样,因为它没有施加在原假定上。

定义非参数残差为

$$\hat{u}_i = Y_i - \hat{m}(X_i) - \bar{\delta},\ i = 1,\cdots,n$$

其中 $\hat{m}(X_i) = \sum_{z=0}^{c-1}\hat{m}(X_i,z)/c$, $\bar{\delta} = \sum_{i=1}^{n}(Y_i - \hat{m}(X_i))/n$ 。以下的简单自举法由 Racine 等(2006)建议。

(i) 令 $u_1^*,\cdots,u_n^*$ 为原始自举法误差,通过 u_i^* 以概率 $r=(1+\sqrt{5})/(2\sqrt{5})$ 取 $u_i^*=[(1-\sqrt{5})/2]\hat{u}_i$,以概率 $1-r$ 取 $u_i^*=[(1+\sqrt{5})/2]\hat{u}_i$ 生成。接着由 $Y_i^* = \hat{m}(X_i)+u_i^*$ 生成 Y_i^*。自举样本为(Y_i^*,X_i,Z_i),$i=1,\cdots,n$,也就是说,$X_i^*=X_i$,$Z_i^*=Z_i$,我们记为$\{Y_i,X_i,Z_i^*\}_{i=1}^n$。

(ii) 利用自举样本来计算自举统计量 I_n^{e*},其中 I_n^{e*} 与 I_n^e 相同,除了以 Z_i^* 取代 Z_i。也就是说,它们使用起初得到的相同交错鉴定选择的 $\hat{h}$、$\hat{\lambda}$ 和 $\hat{\lambda}_z$。

(iii) 重复步骤(i)和(ii)多次,比如 $B=399$ 次。令$\{I_{n,j}^{e*}\}_{j=1}^B$为 B 自举法统计量的有序统计量(升序),令 $I_{n,(\alpha)}^{e*}$ 为$\{I_{n,j}^{e*}\}_{j=1}^B$的 α 分位点。如果在显著水平 α 上有 $I_n^{e*} > I_{n,(\alpha)}^{e*}$则拒绝假定 H_0。

上述自举法的优点在于计算简单。只需要应用一次交错鉴定,而不需要重复计算每一个自举样本的交错鉴定平滑参数。使用原始自举法的另一个优点是,原始自举法不是朴素独立同分布自举法,检验对条件异方差的出现是稳健的。上述过程中 X_i 和 Z_i 都是成对出现的,故它们之间的任何相关可予以保留。 378

Racine(2006)所做的模拟表明了自举法检验表现良好,比传统的基于频率的非参数检验具有显著更好的检验功效。

Lavergne(2001)提出了一个对一组非参数回归函数的等同性检验。该检验也可用于检验一个离散变量的显著性。Hall 和 Hart(1990),King,Hart 和 Wehrly(1991),Delgado(1993),以及 Dette 和 Neumeyer(2001)提出了检验回归曲线

等价的检验统计量。

12.4 概率密度函数的相关检验

12.4.1 两个随机变量间的独立性检验

令$(X,Y)'$为一个$(p+q)\times 1$维随机向量，具有联合累积分布函数$F(x,y)$和联合概率密度函数$f(x,y)$。进一步，令$F_1(x)(F_2(y))$表示$X(Y)$的边缘累积分布函数，有边缘概率密度函数$f_1(x)(f_2(y))$。我们关心的是检验原假定$H_0^f: f(x,y)=f_1(x)f_2(y)$，对几乎所有$(x,y)$。在$H_0^f$假定下，$X$和$Y$相互独立。我们构造一个基于$f(\cdot,\cdot)$与$f_1(\cdot)f_2(\cdot)$积分平方差的检验。由下式给出：

$$
\begin{aligned}
I^f &= \int[f(x,y)-f_1(x)f_2(y)]^2\mathrm{d}x\mathrm{d}y \\
&= \int[f^2(x,y)+f_1^2(x)f_2^2(y)-2f_1(x)f_2(y)f(x,y)]\mathrm{d}x\mathrm{d}y \\
&= \int f(x,y)\mathrm{d}F(x,y)+\int f_1(x)\mathrm{d}F_1(x)\int f_2(y)\mathrm{d}F_2(y) \\
&\quad -2\int f_1(x)f_2(y)\mathrm{d}F(x,y) \\
&= E[f(X,Y)]+E[f_1(X)]E[f_2(Y)]-2E[f_1(X)f_2(Y)]
\end{aligned}
\tag{12.25}
$$

以相应的样本值取代未知的总体均值，以相应的去一核估计值取代式(12.5)中的$f(X_i,Y_i)$、$f_1(X_i)$和$f_2(Y_i)$，可以得到一个可行的检验统计量：

$$
\begin{aligned}
I_n^f &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i,j=1}^{n}K_{h_x,ij}^{x}K_{h_y,ij}^{y}+\frac{1}{n^4}\sum_{i=1}^{n}\sum_{j\neq i,j=1}^{n}K_{h_x,ij}^{x}\sum_{l=1}^{n}\sum_{r\neq l,r=1}^{n}K_{h_y,lr}^{y} \\
&\quad -\frac{2}{n^3}\sum_{i=1}^{n}\sum_{j\neq i,j=1}^{n}\sum_{l\neq i,l=1}^{n}K_{h_x,ij}^{x}K_{h_y,il}^{y}
\end{aligned}
\tag{12.26}
$$

其中，再次有$K_{h_x,ij}^{x}=\prod_{s=1}^{p}h_{x,s}^{-1}k((X_{is}-X_{js})/h_{x,s})$，其中$K_{h_y,ij}^{y}=\prod_{s=1}^{p}h_{y,s}^{-1}k((Y_{is}-Y_{js})/h_{y,s})$。此处我们使用相同的平滑参数集，即$h_{x,s}$、$h_{y,s}$来估计联合概率密度函数$f(x,y)$和边缘概率密度函数$f_1(x)$及$f_2(y)$。

条件 12.5

(i) $f(x,y)$、$f_1(x)$和$f_2(x)$是连续有界函数，$|f_l(u+v)-f_l(u)|\leqslant C|v|$ $(l=1,2)$。

(ii) 单变量核函数k是一个有界、对称、非负的二阶核函数。

(iii) 当$n\to\infty$时，对于$s=1,\cdots,p$，有$h_{x,s}\to 0$；对于$s=1,\cdots,q$有$h_{y,s}\to 0$；且

$nh_{x,1}\cdots h_{x,p}h_{y,1}\cdots h_{y,q}\to\infty$。

Ahmad 和 Li(1997b)证明了以下结论。

定理 12.11 令 $H_{x,p}=h_{x,1}\cdots h_{x,p}, H_{y,q}=h_{y,1}\cdots h_{y,q}$，在条件 12.3 和假定 H_0^f 下，我们有

$$n(H_{x,p}H_{y,q})^{1/2}I_n^f/\hat{\sigma}_f\xrightarrow{d}N(0,1) \tag{12.27}$$

其中 $\hat{\sigma}_f^2=\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(K_{h,ij}^x)^2(K_{h,ij}^y)^2$。

Ahmad 和 Li(1997a)证明了定理 12.11，虽然他们使用了一个不同的方差估计量：

$$\tilde{\sigma}_f^2=2\kappa^{p+q}\left[\frac{1}{n}\sum_{i=1}^{n}\hat{f}_1(X_i)\right]\left[\frac{1}{n}\sum_{i=1}^{n}\hat{f}_2(Y_i)\right]$$

定理 12.11 给出的 $\hat{\sigma}_f^2$ 相对于 $\tilde{\sigma}_f^2$ 在某种程度上更容易计算，因为前者并不要求计算 κ($\kappa=\int k(v)^2\mathrm{d}v$)。

12.4.2 一个参数概率密度函数的检验

380

Fan(1994)考虑了检验一个特定参数概率密度函数的问题。也就是对于几乎所有的 x，检验

$$H_0^g:f(x)=f_0(x,\delta)$$

其中 f_0 是一个未知概率密度函数，δ 是一个未知参数向量。

Fan 考虑了一个基于下式的检验：

$$\begin{aligned}I^g&=\int[f(x)-f_0(x,\delta)]^2\mathrm{d}x\\&=\int[f^2(x)+f_0^2(x,\delta)-2f(x)f_0(x,\delta)]\mathrm{d}x\\&=\int f(x)\mathrm{d}F(x)+\int f_0^2(x,\delta)\mathrm{d}x-2\int f_0(x,\delta)\mathrm{d}F(x)\\&=E[f(X)]+\int f_0^2(x,\delta)\mathrm{d}x-2E[f_0(X,\delta)]\end{aligned}$$

令 $\hat{\delta}$ 表示基于零分布的 δ_0 的极大似然估计值。以 $\hat{f}_{-i}(X_i)$ 和 $\hat{\delta}$ 分别取代 $f(X_i)$ 和 δ 得到一个可行的统计量：

$$I_n^g=\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i)+\int f_0^2(x,\hat{\delta})\mathrm{d}x-\frac{2}{n}\sum_{i=1}^{n}f_0(X_i,\hat{\delta})$$

其中 $\hat{f}_{-i}(X_i)=(n-1)^{-1}\sum_{j\neq i}^{n}K_h(X_j,X_i)$，而 $K_h(X_j,X_i)=\prod_{s=1}^{q}h_s^{-1}k\left(\frac{X_{is}-X_{js}}{h_s}\right)$。

Fan(1994)证明了 I_n^g 统计量有三个不同的渐近分布，取决于数据是否平滑

不足($nh_1\cdots h_q\eta_2^2\to 0,\eta_2=\sum_{s=1}^{q}h_s^2$),最优平滑($nh_1\cdots h_q\eta_2^2\to c>0$),或过度平滑($nh_1\cdots h_q\eta_2^2\to\infty$);也可参见 Pagan 和 Ullah(1999,第 63 页)。

Fan(1994)还提出了偏差修正检验统计量,它在 $h_s\to 0$ 和 $nh_1\cdots h_q\to\infty$ 的一般情况下具有一个简单的渐近分布,因此,数据可以是平滑不足、最优平滑或者过度平滑。我们将在下面给出偏差修正检验。

偏差修正检验需要以核平滑形式 $K_h*f(x,\hat{\delta})=\int K_h(x,v)f_0(v,\hat{\delta})\mathrm{d}v$ 其中

381

的星号($*$)表示卷积,取代 I_n^g 中的 $f_0(x,\hat{\delta})$,这个检验统计量为:

$$I_n^g=n^{-1}\sum_{i=1}^{n}\hat{f}_{-i}(X_i)+\int[K_h*f_0(x,\hat{\delta})]^2\mathrm{d}x-2n^{-1}\sum_{i=1}^{n}K_h*f_0(X_i,\hat{\delta})\tag{12.28}$$

I_n^g 的渐近分布由下一个定理给出。

定理 12.12 在条件 12.5 和 H_0^g 假定条件下,有

$$n(h_1\cdots h_q)^{1/2}I_n^g/\hat{\sigma}_g\xrightarrow{d}N(0,1)$$

其中 $\hat{\sigma}_g^2=\dfrac{2h_1\cdots h_q}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_h(X_i,X_j)^2$。

证明:见 Fan(1994)的定理 4.1。 □

关键是要注意到我们这里给出的检验统计量与 Fan(1994)的定理 4.1 给出的 T_n 统计量有稍许不同。Fan 以 $\int\hat{f}^2(x)\mathrm{d}x$ 取代 $\int f^2(x)\mathrm{d}x$,其中

$$\begin{aligned}\int\hat{f}^2(x)\mathrm{d}x&=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int K_h(X_i,x)K_h(X_j,x)\mathrm{d}x\\&=\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\bar{K}_h(X_i,X_j)\end{aligned}$$

其中 $\bar{K}_h(X_i,X_j)=\prod_{s=1}^{q}h_s^{-1}\bar{k}((X_{is}-X_{js})/h_s)$,且 $\bar{k}(v)=\int k(u)k(v+u)\mathrm{d}u$ 是一个双重卷积核。Fan 的定理 4.2 中的中间项 $(nh_1\cdots h_q)^{-1}\int K(v)^2\mathrm{d}v$ 等于此处的 $(nh_1\cdots h_q)^{-1}\bar{K}(0)$ ($\bar{K}(0)=\bar{k}(0)^q$)。在式(12.28)中,我们以 $\int\hat{f}(x)\mathrm{d}F_n(x)=n^{-1}\sum_{i=1}^{n}\hat{f}_{-i}(X_i)=[n(n-1)]^{-1}\sum_{i=1}^{n}\sum_{j\neq i}^{n}K_h(X_i,X_j)$ 取代 $\int f^2(x)\mathrm{d}x=\int f(x)\mathrm{d}F(x)$。在我们的统计量中,$K(\cdot)$ 与 Fan 的 T_n 统计量中的卷积核 $\bar{K}(\cdot)$ 起的作用相同。因此,如果我们不使用去一核估计量,则我

们的检验有一个中间项等于$(nh_1\cdots h_q)^{-1}k(0)^q$，即$n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}K_h(X_i,X_j)$中的$i=j$项。但由于我们选择使用一个去一核估计值，我们的检验I_n^g不具有渐近非零中间项。总的来说，一个中心值不受限制的检验比一个有非零中心值的检验有更好的有限样本性质。因为后者需要对中心值进行估计，这经常会导致一些有限样本偏差。

12.4.3 条件参数分布的核检验

382

令$\{Y_i,X_i\}_{i=1}^{n}$表示独立同分布数据，其联合概率密度函数由$f(y,x)$给定。令$f(y|x)$表示给定X的Y的条件概率密度函数。我们关注的是检验$f(y|x)$是否属于一个特定的参数族。令$p(y|x,\theta)$表示一个参数条件概率密度函数，θ是一个$k\times 1$维参数向量。原假定由下式给定：

$$H_0^i:P[f(Y_i\mid X_i)=p(Y_i\mid X_i,\theta)]=1,\text{某些 }\theta\in\Theta$$

其中Θ是参数空间的一个紧集。备择假定是对原假定的否定，即对所有$\theta\in\Theta$，$H_1:P[f(Y_i|X_i)=p(Y_i|X_i,\theta)]<1$

Kullback-Leibler 信息准则度量了两个条件累积分布函数之间的差异，该准则由下式定义：

$$I(p,f)=E\left\{\ln\left[\frac{f(Y_i\mid X_i)}{p(Y_i\mid X_i,\theta)}\right]\right\}=\int f(y,x)\ln\left[\frac{f(y\mid x)}{p(y\mid x,\theta_0)}\right]\mathrm{d}y\mathrm{d}x \tag{12.29}$$

我们都知道$I(p,f)\geqslant 0$，当且仅当$f(y|x)=p(y|x,\theta_0)$几乎处处有$I(p,f)=0$。因此，$I(p,f)$可以作为检验H_0成立与否的合适度量。Zheng(2000)考虑了信息方程的一阶展开式，即

$$\begin{aligned}E\left\{\ln\left[\frac{f(Y_i\mid X_i)}{p(Y_i\mid X_i,\theta)}\right]\right\}&\cong E\left[\frac{f(Y_i\mid X_i)}{p(Y_i\mid X_i,\theta)}-1\right]\\&=E\left[\frac{f(Y_i\mid X_i)-p(Y_i\mid X_i,\theta)}{p(Y_i\mid X_i,\theta)}\right]\end{aligned} \tag{12.30}$$

以密度$f_1(x)$对式(12.30)进行加权得到

$$I_1(p,f)=E\left[\frac{f(Y_i\mid X_i)f_1(X_i)-p(Y_i\mid X_i,\theta)f_1(X_i)}{p(Y_i\mid X_i,\theta)}\right] \tag{12.31}$$

Zheng(2000)证明了$I_1(f,p)\geqslant 0$，当且仅当H_0成立时等号成立。因此，$I_1(f,p)$也可以看成是检验H_0成立与否的合适度量。Zheng 考虑了X和Y都是连续变量的情况。Fan 等(2006)将 Zheng 的检验扩展到了离散因变量和混合条件协变量的情况。下面我们首先讨论因变量Y离散的情况(X包含混合数据类型)。这包括了很多有趣的情况，如二元或多项式响应模型参数的正确设定检

383

验。令 $X=(X^c,X^d)$，其中 X^c 为 $q\times1$ 维连续变量，X^d 为 $r\times1$ 维离散变量。我们用 $X_{is}^c(X_{is}^d)$ 表示 $X^c(X^d)$ 的第 s 个元素。令 $l(X_{is}^d,X_{js}^d,\lambda_s)$ 表示在第 4 章介绍的关于离散变量 X_s^d 的核，而离散变量的乘积核为 $L(X_i^d,X_j^d,\lambda)=\prod_{s=1}^{r} l(X_{is}^d,X_{js}^d,\lambda_s)$。对于连续元素 X^c，我们仍旧采用之前讨论的标准（二阶）乘积核函数。因此，对于混合数据变量 $X=(X^c,X^d)$，核函数由式(12.32)定义（$\gamma=(h,\lambda)$）

$$K_{\gamma,ij}\overset{\text{def}}{=}W_h\left(\frac{X_i^c-X_j^c}{h}\right)L(X_i^d,X_j^d,\lambda)\tag{12.32}$$

当 Y_i 是一个离散变量时，我们通过下面的去一核估计值来估计 $f(Y_i|X_i)$ 和 $f_1(X_i)$（我们没有平滑 y）：

$$\hat{f}(Y_i,X_i)=\frac{1}{n}\sum_{j\neq i}^{n}\mathbf{1}(Y_i=Y_j)K_n(X_i,X_j)\tag{12.33}$$

$$\hat{f}_1(X_i)=\frac{1}{n}\sum_{j\neq i}^{n}K_n(X_i,X_j)\tag{12.34}$$

我们通过下式给定的一个 $p(\cdot|\cdot,\hat{\theta}_0)$ 的核加权形式来估计 $p(Y_i|X_i,\theta_0)$。

$$\tilde{f}(Y_i,X_i)=\frac{1}{n}\sum_{j=1}^{n}K_{\gamma,ij}p(Y_j|X_j,\hat{\theta})\tag{12.35}$$

其中 $\hat{\theta}$ 是基于原模型的 θ_0 的最大似然估计值。

由上述定义的 $\hat{f}(Y_i,X_i)$、$\hat{f}_1(X_i)$ 和 $\tilde{f}(Y_i,X_i)$，可以得到检验统计量：

$$I_{n,d}^i=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\left\{\frac{K_{\gamma,ij}}{p(Y_i|X_i,\hat{\theta})}[\mathbf{1}(Y_i=Y_j)-p(Y_j|X_j,\hat{\theta})]\right\}\tag{12.36}$$

Fan 等(2006)推荐使用第 5 章中讨论的交错鉴定过程来选择平滑参数 $h_1,\cdots,h_q,\lambda_1,\cdots,\lambda_r$。$I_{n,d}^i$ 的渐近分布将在下一个定理给出。

384

定理 12.13 在 Fan 等(2006)给出的条件下，在 H_0^i 下我们有

$$T_{n,d}^i\overset{\text{def}}{=}n(h_1\cdots h_q)^{1/2}I_{n,d}^i/\hat{\sigma}_i\to N(0,1)\text{ 依分布收敛}$$

其中

$$\hat{\sigma}_i^2=\frac{2(h_1\cdots h_q)}{n(n-1)}\sum_i\sum_{i\neq j}K_n(X_i,X_j)^2\left\{\frac{[\mathbf{1}(Y_i=Y_j)-p(Y_i|X_j,\hat{\theta})]}{\hat{p}(Y_i|X_i,\hat{\theta})}\right\}^2$$

是 $\sigma_0^2=[\int W^2(v)\mathrm{d}v]E[(1-p_i)p_i^{-1}f_1(X_i)]$ 的一个一致估计量，也是 $n(h_1\cdots h_q)^{1/2}I_{n,d}^i$ 的渐近方差。

可以证明在 H_1^i 下，$T_{n,d}^i$ 发散至 $+\infty$，因此它是一个一致性检验。Fan 等

(2006)中的模拟证明如果我们使用渐近标准正态临界值，$T^{i}_{n,d}$ 检验会遭受有限样本尺度的扭曲。Fan 等建议使用以下的自举法程序来更好地逼近 $T^{i}_{n,d}$ 有限样本原分布。

(i) 基于参数条件分布 $p(Y_i^* | X_i, \hat{\theta})$ 生成 Y_i^*。称 $\{X_i, Y_i^*\}_{i=1}^{n}$ 为自举样本。

(ii) 基于参数原模型，使用自举样本估计 θ。令 $\hat{\theta}^*$ 代表得到的估计量。用与 $T^{i}_{n,d}$ 相同的方法计算自举统计量 $T^{i*}_{n,d}$，除了用 Y_i^* 和 $\hat{\theta}^*$ 分别替代 Y_i 和 $\hat{\theta}$。

(iii) 重复步骤(i)和(ii)多次，比如 $B=399$，得到 $\{T^{i*}_{n,d}\}_{j=1}^{B}$ 的经验分布。令 $T^{i*}_{n,d}$ 代表自举分布的 α 分位点。如果 $T^{i}_{n,d} > T^{i*}_{n,\alpha}$，我们在显著水平 α 上拒绝原假定。

接下来我们将讨论因变量 Y 是连续的情形。为了解释的简单化，我们仅考虑 Y 是一个标量的情形。Zheng(2000)建议使用一个平滑密度估计量 $\tilde{f}(Y_i, X_i)$ 来估计 $p(Y_i | X_i, \theta_0) f_1(X_i)$，其中

$$\tilde{f}(Y_i, X_i) = \frac{1}{n} \sum_{j=1}^{n} \int W_{2,h_y}\left(\frac{Y_i - y}{h_y}\right) K_h\left(\frac{X_i - X_j}{h}\right) f(y \mid X_j, \hat{\theta}) \mathrm{d}y \qquad (12.37)$$

其中 $W_{2,h_y}(\cdot) = h_y^{-1} W_2(\cdot)$，$W_2(\cdot)$(特别定义地)是一个单变量的核函数。Zheng 建议首先将因变量 Y 转换使其落在 $[0,1]$ 之间，然后为 $W_2(\cdot)$ 选择一个特殊的核函数，这个核函数拥有以下性质，即当 $n \to \infty$ 时，$h_y^{-1}\int_0^1 W_2((Y_i - y)/h_y)^2 \mathrm{d}y \to 1$。Zheng 提出下面的检验统计量：

385

$$I^{i}_{n,c} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left[\frac{W_{2,h_y}\left(\frac{Y_i - Y_j}{h_y}\right) K_{\gamma,ij} - \int W_{2,h_y}\left(\frac{Y_i - y}{h_y}\right) K_{\gamma,ij} f(y \mid X_j, \hat{\theta}) \mathrm{d}y}{f(Y_i \mid X_i, \hat{\theta})}\right] \qquad (12.38)$$

$I^{i}_{n,c}$ 的渐近分布由下面的定理给出。

定理 12.14　在 Fan 等(2006)给出的条件下，在 H_0^i 假定下，我们有

$$T^{i}_{n,c} \stackrel{\text{def}}{=} n(h_y h_1 \cdots h_q)^{1/2} I^{i}_{n,c} / \hat{\sigma}_{i,c} \to N(0,1) \text{ 依分布收敛}$$

其中

$$\hat{\sigma}^2_{i,c} = \frac{2(h_y h_1 \cdots h_q)}{n(n-1)} \sum_{i=1}^{n} \sum_{j\neq i}^{n} K^2_{\gamma,ij}$$

其他的模型设定检验包括 Gozalo 和 Linton(2001)以及 Härdle, Sperlich 和 Spokoiny(2001)，他们提出了一个基于核的检验，用来检验一般的非参数回归模型的可加性。Sun(2006)提出了一种检验条件分位数函数是否等同的检验。

12.5 应 用

12.5.1 增长收敛俱乐部

Quah(1996),Maasoumi 等(2007)和其他学者对是否存在“收敛俱乐部”的观点进行了检验,也就是说,对于俱乐部,如经济合作发展组织(OECD),成员的增长率是否不同。我们并不打算在这里回顾大量的文献,而是建议感兴趣的读者参见 Mankiw,Romer 和 Weil(1992),Liu 和 Stengos(1999),Durlauf 和 Quah(1999)以及其中的参考文献。接下来,我们沿用 Maasoumi 等的分析。

386

我们应用式(12.22)中的检验统计量来判定 OECD 国家和非 OECD 国家是否遵从同样的增长模型。我们通过检验 OECD 成员(一个二元分类变量)在非参数框架下是否是一个相关的回归元做到这一点。原假定为 OECD 成员是一个非相关的回归元,所以在原假定的前提下,OECD 和非 OECD 国家的增长率是由相同的增长模型决定的。备择假定是原假定的反面,例如,OECD 和非 OECD 国家有不同增长率(回归)模型。

按照 Liu 和 Stengos(1999)的方法,我们利用 88 个国家在过去七个五年(1960—1964,1965—1969,1970—1974,1975—1979,1980—1984,1985—1989,1990—1994)的面板数据来得到面板中的 $88\times7=616$ 个观测值。我们通过下面的模型来构建我们的检验:

$$\text{Growth}_{it}=m(\text{OECD}_{it},\text{dT}_t,\ln(\text{inv}_{it}),\ln(\text{popgro}_{it}),\ln(\text{inigdp}_{it}),\ln(\text{humancap}_{it}))+\varepsilon_{it} \tag{12.39}$$

其中 Growth_{it}指每期的人均收入的增长率,dT_t 是七个时期的虚拟变量,inv_{it}是指投资占国内生产总值(GDP)的比率, popgro_{it}代表劳动力的增长, inigdp_{it}代表每期开始的人均收入, humancap_{it}指人力资本。初始收入估计来自 Summers-Heston 的数据库,五年期的平均投资/GDP 比率的估计也是一样。人均 GDP 的平均增长率和每期的平均每年人口增长率是从世界银行得到的。最后,超过 15 岁的平均年受教育水平是从 Barro 和 Lee(2000)的文章中得到的。

在我们对基于平滑的非参数检验的结果进行报告之前,我们首先考虑一些应对此问题流行的参数方法。一种常用的参数方法是应用一个线性回归模型,把 OECD 虚拟变量作为一个可行的回归元,然后检验这个虚拟变量的系数是否显著。我们考虑 Liu 和 Stengos(1999)提出的一个参数设定,其中包含 OECD 国家状态和年份的虚拟变量,初始的 GDP 和人力资本变量是非线性的,由下式给出:

387

$$\text{Growth}_{it} = \beta_0 \text{OECD}_{it} + \sum_{s=1}^{7} \beta_s \text{dT}_s + \beta_8 \ln(\text{inv}_{it}) + \beta_9 \ln(\text{popgro}_{it}) + \sum_{s=1}^{4} \alpha_s [\ln(\text{inigdp}_{it})]^s + \sum_{s=1}^{3} \gamma_s [\ln(\text{humancap})_{it})]^s + \varepsilon_{it} \quad (12.40)$$

模型(12.40)的估计结果得到 OECD 虚拟变量的 t 统计量等于 -0.973,P 值为 0.33。于是,参数检验不能拒绝原假定。

接下来,我们遵循传统的频率方法并使用非参数检验,即我们的检验是基于模型(12.39),其中样本根据 OECD 和 dT 虚拟变量划分。使用 $B=999$ 次自举再抽样,我们得到 0.113 的 P 值,我们以传统的 1%、5%、10% 的水平再一次拒绝原假定。

我们现在报告以平滑为基础的非参数检验的结果(见 12.3.5 节)。对于每一个自举检验,我们采用 $B=999$ 次自举再抽样,而对于交错鉴定检验,我们应用 5 个重新开始的数值搜索算法,并保留那些产生交错鉴定函数的最低值的平滑参数。在 I_n^a 下从经验累积分布函数的反函数得到的 P 值为 0.006,这个值是否决原假定有效性的强有力的证据。

参数检验和非参数检验的不一致性也显示了参数模型的设定是错误的。所以对于参数模型的正确设定我们应用一致性非参数检验 I_n^a(见 12.1.1 节)。这个检验的 P 值为 0.001,因此我们拒绝正确参数设定的原假定。

传统的基于频率的非参数检验同样不能拒绝原假定的原因是当我们估计非参数回归函数时,它把样本分成 $2\times7=14$ 个部分(离散单元数来自离散变量 OECD 和 dT),所以对于传统的基于频率的检验,较小的(子)样本容量导致相当大的有限样本功效损失。

我们得出结论,稳健的非参数证据支持收敛俱乐部的存在,这是一个当使用一般的参数估计和传统的非参数方法仍然没有检测到的特征。也就是说,OECD 国家的增长率相比非 OECD 国家是通过不同的增长模式而产生的。 388

12.6 证　明

12.6.1 定理 12.1 的证明

证明:我们将在 $m(X_i,\gamma)=\alpha+X_i'\beta\equiv Z_i'\gamma$ 的情况下证明定理 12.1,其中 $Z_i=(1,X_i')$,$\gamma=(\alpha,\beta')'$。一般情况的证明过程类似,将留作习题(参见习题 12.4)。使用 $\hat{u}_i=Y_i-Z_i'\hat{\gamma}=u_i-Z_i'(\hat{\gamma}-\gamma)$,可得(运用 $n(n-1)\sim n^2$)。

$$
\begin{aligned}
I_n^e &= n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} u_i u_j K_{h,ij} - 2(\hat{\gamma}-\gamma)' n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} u_i Z_j K_{h,ij} \\
&\quad + (\hat{\gamma}-\gamma)' n^{-2}\sum_{i=1}^{n}\sum_{j\neq i}^{n} Z_i Z_j' K_{h,ij}(\hat{\gamma}-\gamma) \\
&\equiv I_{1n}^a - (\hat{\gamma}-\gamma)' I_{2n}^a + (\hat{\gamma}-\gamma)' I_{3n}^a (\hat{\gamma}-\gamma)
\end{aligned}
$$

其中 I_{jn}^a 的定义是显而易见的($j=1,2,3$)。

记 $I_{1n}^a = 2n^{-2}\sum_{i=1}^{m}\sum_{j>i}^{n} u_i u_j K_{h,ij}$,可以很容易证明

$$
\begin{aligned}
E[(I_{1n}^a)^2] &= 4n^{-4}\sum_{i}\sum_{j>i} E[u_i^2 u_j^2 K_{h,ij}^2] \\
&= \frac{2}{n^2 H_q^2}\{E[\sigma^2(X_1)\sigma^2(X_2)K((X_1-X_2)/h)^2] + o(1)\} \\
&= \frac{2}{n^2 H_q^2}\left\{\int f(x_1)f(x_2)\sigma^2(x_1)\sigma^2(x_2)K((x_1-x_2)/h)^2 \mathrm{d}x_1 \mathrm{d}x_2 + o(1)\right\} \\
&= \frac{2}{n^2 H_q}\left\{\int f(x_1)f(x_1+hv)\sigma^2(x_1)\sigma^2(x_1+hv)K(v)^2 \mathrm{d}x_1 \mathrm{d}v + o(1)\right\} \\
&= \frac{2}{n^2 H_q}\left\{k^q \int f^2(x_1)\sigma^4(x_1)\mathrm{d}x_1 + o(1)\right\} \\
&= \frac{2}{n^2 H_q}\{k^q E[f(X_1)\sigma^4(X_1)] + o(1)\} \equiv (n^2 H_q)^{-1}\{\sigma_a^2 + o(1)\}
\end{aligned}
$$

因此,$nH_q^{1/2}I_{1n}^a$ 具有均值 0 和渐近方差 σ_a^2。由 Hall(1984)的中心极限定理(见附录 A 引理 A.16),我们知道

$$
nH_q^{1/2}I_{1n}^a \xrightarrow{d} N(0,\sigma_a^2)
$$

389 令 $I_{2n,t}^a, I_{3n,ts}^a$ 分别表示 I_{2n}^a, I_{3n}^a 的第 t 个和第 (t,s) 个元素。很容易证明,对于所有的 t 有 $E[|I_{2n,t}^a|^2] = O(n^{-1})$,这意味着 $I_{2n}^a = O_p(n^{-1/2})$;同样,对于所有的 t、s 有 $E[|I_{3n,ts}^a|] = O(1)$,这意味着 $I_{3n}^a = O_p(1)$。这些结果与 $\hat{\gamma}-\gamma = O_p(n^{-1/2})$ 一起可推出

$$
nH_q^{1/2}I_n^a = nH_q^{1/2}I_{1n}^a + O_p(H_q^{1/2}) \xrightarrow{d} N(0,\sigma_a^2) \tag{12.41}
$$

习题 12.3 要求读者证明 $\hat{\sigma}_a^2 = \sigma_a^2 + o_p(1)$。这就完成了对定义 12.1 的证明。 □

12.6.2 定理 12.2 的证明

证明:我们只考虑当 $m(X_i,\gamma) = \alpha + X_i'\beta = Z_i'\gamma$ 时的情况。令 $g(X_i) = E(Y_i|X_i)$,令 $\bar{\gamma}$ 表示 H_1 假定条件下 $\hat{\gamma}$ 的概率极限。类似于定理 12.1 的证明,

我们可以证明 H_1 假定条件下，$I_n^a = I_{n,0} + o_p(1)$，其中

$$I_{n,0} = \frac{1}{n^2}\sum_i \sum_{j\neq i} [u_i + g(X_i) - Z'_i\bar{\gamma}][u_j + g(X_j) - Z'_j\bar{\gamma}]K_h(X_i - X_j)$$

习题 12.4 要求读者证明 $I_{n,0} = C + o_p(1)$，其中 $C = E[f(X_i)(g(X_i) - Z_i'\bar{\gamma})^2] > 0$。因此，$nH_q^{1/2}I_n^a = nH_q^{1/2}C + o_p(nH_q^{1/2})$，这暗含了定理 12.2。□

12.6.3　定理 12.5 的证明

证明：检验统计量 I_n^b 可以写成 $I_n^b = I_{1n}^b + I_{2n}^b$，其中

$$I_{1n}^b = c_{n,b} - 2(n_1, n_2)^{-1}\sum_{i=1}^{n} K_{h,ii}^{xy}$$

其中 $c_{n,b} = k(0)^q(h_1\cdots h_q)^{-1}[n_1^{-1} + n_2^{-1}]$，$n = \min\{n_1, n_2\}$，且

$$I_{2n}^b = \sum_i \sum_{j\neq i}\left[\frac{1}{n_1^2}K_{h,ij}^x + \frac{1}{n_2^2}K_{h,ij}^y - \frac{1}{n_1 n_2}K_{h,ij}^{xy} - \frac{1}{n_1 n_2}K_{h,ij}^{yx}\right]$$

其中如果被加数包含 X_i，则 $\sum_i = \sum_{i=1}^{n_1}$；如果被加数包含 Y_i，则 $\sum_i = \sum_{i=1}^{n_2}$，$\sum_j$ 的定义类似。

定义 $H_q = h_1\cdots h_q$，很容易证明

$$E[I_{1n}^b] = c_{n,b} + (n_1, n_2)^{-1}O(nH_q) = c_{n,b}[1 + O(n^{-1}H_q)]$$

且　390

$$\mathrm{var}(I_{1n}^b) = (n_1 n_2 H_q)^{-2}O(n^2H_q^2 + nH_q) = O(n^{-2})$$

（见习题 12.4）因此，

$$I_{1n}^b = c_n, b + O_p(n^{-1}) \tag{12.42}$$

令 $z_i = (X_i, Y_i)$，定义 $H_n(z_i, z_j) = K_{h,ij}^x + K_{h,ij}^y - K_{ij}^{h,xy} - K_{ij}^{h,yx}$。对于 $i\neq j$，在 H_0^b 假定条件下（由于 $f = g$）有

$$E[H_n(z_i, z_j) \mid z_i] = \{E[k_{h,ij}^x \mid X_i] - E[k_{h,ij}^{xy} \mid X_i]\} + \{E[k_{h,ij}^y \mid Y_i] - E[k_{h,ij}^{yx} \mid Y_i]\} = 0$$

因此，I_{1n}^b是一个退化的 U 统计量（见附录 A 中的定义）。很容易看出 $E[(I_{2n}^b)^2] = \mathrm{var}(I_{2n}^b) = O((n_1 n_2 H_q)^{-1})$，且

$$\begin{aligned}
\mathrm{var}[(n_1 n_2 H_q)^{1/2}I_{2n}^b] &= E\{[(n_1 n_2 H_q)^{1/2}I_{2n}^b]^2\} \\
&= (n_1 n_2 H_q)\sum_i \sum_{j\neq i}\{n_1^{-4}E[(k_{h,ij}^x)^2] + n_2^{-4}E[(k_{h,ij}^y)^2] \\
&\quad + (n_1 n_2)^{-2}E[(k_{h,ij}^{x,y})^2] + (n_1 n_2)^{-2}E[(k_{h,ij}^{y,x})^2] + O(n^{-4})\} \\
&\sim (n_1 n_2 H_q)\sum_i \sum_{j\neq i}\{n_1^{-4}(k_{h,ij}^x)^2 + n_2^{-4}(k_{h,ij}^y)^2
\end{aligned}$$

$$+2(n_1n_2)^{-2}(k_{h,ij}^{x,y})^2\}$$

$$\equiv \tilde{\sigma}_a^2 = O_p(1)$$

上述 $A \sim B$ 意味着 $A = B + o_p(1)$。因此，$\tilde{\sigma}_a^2$ 是 $E[(n_1n_2H_q)^{1/2}I_{2n}^b]^2$ 的一个一致估计量。注意

$$\hat{\sigma}_a^2 - \tilde{\sigma}_a^2 = \frac{n_1n_2}{h_1\cdots h_q}\sum_{i=1}^{n}\left\{\frac{k(0)^{2q}}{n_1^4} + \frac{k(0)^{2q}}{n_2^4} + \frac{2h_1\cdots h_q}{(n_1n_2)^2}k_h(X_i,Y_i)^2\right\}$$

$$= O_p((nh_1\cdots h_q)^{-1}) = o_p(1)$$

因此，$\hat{\sigma}_a^2$ 是 $\operatorname{var}[(n_1n_2H_q)^{1/2}I_{2n}^b]$ 的一个一致估计量。

最后，关于退化 U 统计量的成立（见附录 A），可直接检查 Hall(1984) 中心极限定理的条件。因此，我们有

$$(n_1n_2H_q)^{1/2}I_{2n}^b/\hat{\sigma}_b \xrightarrow{d} N(0,1) \tag{12.43}$$

391

式(12.42)和式(12.43)可以推出

$$(n_1n_2H_q)^{1/2}[I_n^b - c_{n,b}]/\hat{\sigma}_b = (n_1n_2H_q)^{1/2}I_{2n}^b/\hat{\sigma}_b + o_p(1) \xrightarrow{d} N(0,1)$$

这就完成了定理 12.5 的证明。 □

12.6.4 定理 12.9 的证明

令 $r_i = r(W_i) \equiv E(Y_i|W_i)$。然后使用 $\tilde{U}_i = Y_i - \hat{Y}_i = (r_i - \hat{r}_i) + u_i - \hat{u}_i$，其中 $\hat{r}_i = n^{-1}\sum_{j\neq i} r_jL_{h_w,ij}/\hat{f}_{w_i}$，$\hat{u}_i = n^{-1}\sum_{j\neq i} u_jL_{h_w,ij}/\hat{f}_{w_i}$，我们有如下 I_n^d 的表达式：

$$I_n^d = \frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\{(r_i - \hat{r}_i)(r_j - \hat{r}_j)$$

$$+ u_iu_j + \hat{u}_i\hat{u}_j + 2u_i(r_j - \hat{r}_j) - 2\hat{u}_i(r_j - \hat{r}_j) - 2u_i\hat{u}_j\}K_{h,ij}$$

$$\stackrel{\text{def}}{=} I_{n1}^d + I_{n2}^d + I_{n3}^d + 2I_{n4}^d - 2I_{n5}^d - 2I_{n6}^d$$

为了记号的简洁，在下面的证明过程中我们假定 $h_1 = \cdots = h_q, h_{w,1} = \cdots = h_{w,q} = h_w$。我们将通过证明 $I_{ni}^d = o_p((nh^{q/2})^{-1})$，$i = 1,3,4,5,6$，以及在分布上 $nh^{q/2}I_{n2}^d/\hat{\sigma}_d \to N(0,1)$ 来完成定理 12.9 的证明。在以下的引理 12.1 至引理12.6 中这些结论将被证明。

引理 12.1 $J_{n1}^d = o_p((nh^{q/2})^{-1})$。

证明：注意 $K(\cdot)$ 非负，且 $\hat{f}_{X_i} = n^{-1}\sum_{j\neq i} K_{h,ij}(n-1 \sim n)$，我们有

$$I_{n1}^d = n^{-2}\sum_i\sum_{j\neq i}(r_i - \hat{r}_i)(r_j - \hat{r}_j)K_{h,ij}$$

$$\leqslant (1/2)n^{-2}\sum_i\sum_{j\neq i}[(r_i - \hat{r}_i)^2 + (r_j - \hat{r}_j)^2]K_{h,ij}$$

$$= n^{-2}\sum_i \sum_{j\neq i} (r_i - \hat{r}_i)^2 K_{h,ij} = n^{-1}\sum_i (r_i - \hat{r}_i)^2 \hat{f}_{X_i}$$

$$\sim n^{-1}\sum_i (r_i - \hat{r}_i)^2 f_{X_i} = O_p(h_w^{2\nu} + (nh_w^p)^{-1})$$

$$= o((nh^{q/2})^{-1})$$

倒数第二行由第 7 章引理 7.2 得到。 □

392

引理 12.2

(i) $nh^{q/2}I_{n2}^d \to N(0,\sigma_d^2)$ 依分布收敛。

(ii) $\hat{\sigma}_d^2 = \sigma_d^2 + o_p(1)$。

(i) 由定理 12.1 证明,(ii) 可以用引理 A.16 容易地证明。

引理 12.3　$I_{n3}^d = o_p((nh^{q/2})^{-1})$。

证明:

$$J_{n3} = n^{-2}\sum_i \sum_{j\neq i} \hat{u}_i \hat{u}_j K_{h,ij}$$

$$\leqslant (1/2)n^{-2}\sum_i \sum_{j\neq i} [\hat{u}_i^2 + \hat{u}_j^2] K_{h,ij}$$

$$= (n^2 h^d)^{-1}\sum_i \sum_{j\neq i} \hat{u}_i^2 K_{h,ij}$$

$$= n^{-1}\sum_i \hat{u}_i^2 \hat{f}_{X_i}$$

$$= n^{-1}\sum_i \hat{u}_i^2 f_{X_i} + n^{-1}\sum_i \hat{u}_i^2 (\hat{f}_i - f_i)$$

$$= [O(1) + o_p(1)] O((nh_w^p)^{-1})$$

$$= o((nh^{q/2})^{-1})$$

由第 7 章引理 7.2 可得。 □

引理 12.4　$I_{n4}^d = o_p((nh^{q/2})^{-1})$。

393

证明:使用恒等式 $\hat{f}_j^{-1} = f_j^{-1}[1 + (f_j - \hat{f}_j)/\hat{f}_j]$,据引理 7.5 我们有

$$I_{n4}^d = n^{-2}\sum_i \sum_{j\neq i} u_i (r_j - \hat{r}_j) K_{h,ij}$$

$$= (n^2 h^d)^{-1}\sum_i \sum_{j\neq i} u_i (r_j - \hat{r}_j)\hat{f}_j K_{h,ij} \hat{f}_j^{-1}$$

$$\sim n^{-2}\sum_i \sum_{j\neq i} u_i (r_j - \hat{r}_j)\hat{f}_j K_{h,ij} f_j^{-1}$$

$$\equiv I_{n4,1}^d$$

$$I_{n4,1}^d = n^{-2}\sum_i \sum_{j\neq i} u_i (r_j - \hat{r}_j)\hat{f}_j K_{h,ij} f_j^{-1}$$

$$= O_p(h_w^{2\nu} + (nh_w^p)^{-1})$$

$$= o_p((nh^{q/2})^{-1})$$

□

引理 12.5 $I_{n5}^{d}=o_p((nh^{q/2})^{-1})$。

证明：由引理 12.1 和 12.3 的证明，

$$\begin{aligned}I_{n5}^{d} &= n^{-2}\sum_{i}\sum_{j\neq i}\hat{u}_i(r_j-\hat{r}_j)\\&\leqslant (1/2)(n^2h^d)^{-1}\sum_{i}\sum_{j\neq i}[(r_j-\hat{r}_j)^2K_{h,ij}+\hat{u}_i^2K_{h,ij}]\\&= O_p(h_w^{2\nu}+(nh_w^p)^{-1})\\&= o((nh^{q/2})^{-1})\end{aligned}$$

□

引理 12.6 $I_{n6}^{d}=o_p((nh^{q/2})^{-1})$。

394 证明：由引理 7.5 得

$$\begin{aligned}I_{n6}^{d} &= n^{-2}\sum_{i}\sum_{j\neq i}u_i\hat{u}_jK_{h,ij}\\&= (n^2h^d)^{-1}\sum_{i}\sum_{j\neq i}u_i\hat{u}_j\hat{f}_jK_{h,ij}\hat{f}_j^{-1}\\&= n^{-2}\sum_{i}\sum_{j\neq i}u_i\hat{u}_j\hat{f}_jK_{h,ij}f_j^{-1}[1+(\hat{f}_j-f_j)/\hat{f}_j]\\&\sim n^{-2}\sum_{i}\sum_{j\neq i}u_i\hat{u}_j\hat{f}_jK_{h,ij}f_j^{-1}\equiv I_{n6,1}^{d}\end{aligned}$$

$$\begin{aligned}I_{n6,1}^{d} &= n^{-2}\sum_{i}\sum_{j\neq i}u_i\hat{u}_j\hat{f}_jK_{h,ij}f_j^{-1}\\&= O_p((nh_wp)^{-1})\\&= o((nh^{q/2})^{-1})\end{aligned}$$

□

12.7 习　题

习题 12.1 假定式(12.4)是真实模型，但是人们却针对模型(12.3)来检验模型(12.2)。同时假定 $E(X_i^3)=E(X_i^5)=E(X_i)=0$。

(i) 证明 $\hat{\beta}_2\xrightarrow{p}0$，其中 $\hat{\beta}_2$ 是基于模型(12.3)β_2 的最小二乘估计值。

(ii) 证明标准 t 检验不具有渐近趋于 1 的功效(即不是一个一致性检验)。

习题 12.2 在 12.1.1 节讨论的原始自举法过程中，证明 $E^*(u_i^*)=0$，$E^*(u_i^{*j})=\hat{u}_i^j, j=2,3$。其中 $E^*(\cdot)=E(\cdot\mid\{X_i,Y_i\}_{i=1}^n)$。

395 **习题 12.3** 假定 $g(X_i,\gamma)=\alpha+X_i'\beta\equiv Z_i'\gamma$，其中，$Z_i'=(1,X_i')$，$\gamma=(\alpha,\beta')'$。由 $\hat{u}_i=u_i-Z_i'(\hat{\gamma}-\gamma)$，我们有

$$I_n^a=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}u_iu_jK_{h,ij}-\frac{2(\hat{\gamma}-\gamma)'}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}u_iZ_jK_{h,ij}$$

$$+(\hat{\gamma}-\gamma)'\left[\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}Z_iZ'_jK_{h,ij}\right](\hat{\gamma}-\gamma)$$

$$\equiv I_{1n}-2(\hat{\gamma}-\gamma)I_{2n}+(\hat{\gamma}-\gamma)'I_{3n}(\hat{\gamma}-\gamma)$$

其中 I_{jn} 的定义是显而易见的。

(i) 证明 $I_{2n}=O_p(n^{-1/2})$, $I_{3n}=O_p(1)$。

提示:先证明对于所有的 $t,s=1,\cdots,q+1$ 有 $E[|I_{2n}|^2]=O(n^{-1})$, $E[|I_{3n,ts}|]=O(1)$,其中 $I_{3n,ts}$ 是 I_{3n} 第 (t,s) 个元素(I_{3n} 为 $(q+1)\times(q+1)$ 维)。

(ii) 证明 $\hat{\sigma}_a^2=\sigma_a^2+o_p(1)$。

提示:证明 $E[\hat{\sigma}_a^2]=\sigma_a^2+o(1)$,并运用 H 分解结果(引理 A.15)。

(iii) 证明 $nH_q^{1/2}I_{1n}\xrightarrow{d}N(0,\sigma_a^2)$。

提示:首先证明 $E(I_{1n})=0$, $\mathrm{var}(nH_q^{1/2}I_{1n})=\sigma_a^2+o(1)$。对于 I_{1n} 检验,验证 Hall(1984)的定理1给出的条件成立,即定义 $H_n(W_i,W_j)=u_iu_jK((X_i-X_j)/h)$, $W_i=(u_i,X_i)$, $G_n(W_i,W_j)=E[H_n(W_i,W_l)H_n(W_j,W_l)\mid W_i,W_j]$。证明

$$\frac{E[G_n^2(W_1,W_2)]+n^{-1}E[H_n^4(W_1,W_2)]}{\{E[H_n^2(W_1,W_2)]\}^2}$$

$$=\frac{O(H_q^3)+n^{-1}O(H_q)}{O(H_q^2)}$$

$$=O(H_q+(nH_q)^{-1})=o(1)$$

(iv) 根据(i)和(ii)的结果证明 $nH_q^{1/2}I_n^a/\sqrt{\hat{\sigma}_a^2}\xrightarrow{d}N(0,1)$。

习题 12.4 类似于定理12.2证明过程中的定义,令

$$I_{n,0}=n^{-2}\sum_i\sum_{j\neq i}[u_i+g(X_i)-Z'_i\bar{\gamma}][u_j+g(X_j)-Z'_j\bar{\gamma}]K_h(X_i,X_j)$$

(i) 证明 $E[I_{n,0}]=E[(g(X_i)-Z_i'\bar{\gamma})^2]+o(1)$。

396

(ii) 证明 $I_{n,0}=E[(g(X_i)-Z_i'\bar{\gamma})^2]+o_p(1)$。

提示:注意 $I_{n,0}$ 可以写成 U 统计量的二阶形式,运用 U 统计量的 H-分解(见附录A),根据(i)很快得到(ii)。

习题 12.5 类似于12.1.1节关于 $\hat{\gamma}^*$ 的定义,证明 $\hat{\gamma}^*-\gamma=O_p(n^{-1/2})$。

提示:记 $\hat{\gamma}^*-\gamma=\hat{\gamma}^*-\hat{\gamma}+\hat{\gamma}-\gamma$。证明 $E^*[(\hat{\gamma}^*-\hat{\gamma})^2]=O_p(n^{-1})$。

习题 12.6 根据定理12.5定义的 I_n^b、$\hat{\sigma}_b^2$ 及 σ_b^2,考虑当 $n_1=n_2=n$ 时的情形。

(i) 证明 I_n^b 是一个退化 U 统计量,也就是证明 $E[K_{h,ij}^x+K_{h,ij}^y-2K_{h,ij}^{x,y}-K_{h,ij}^{y,x}\mid X_i,Y_i]=0$。

(ii) 证明 $\mathrm{var}(n(h_1\cdots h_q)^{1/2}I_n^b)=\sigma_b^2+o(1)$。

(iii) 证明 $\hat{\sigma}_b^2 = \sigma_b^2 + o_p(1)$。

(iv) 运用上面(i)—(iii)的结论证明定理12.5。

提示:在(ii)中记 $\sum_i \sum_{j \neq i} = 2 \sum_i \sum_{j < i}$,并运用(i)的结果。运用引理A.16证明(iii)。对(iv)检查Hall(1984)的中心极限条件。

习题12.7 证明 $\hat{\sigma}_f^2 = \dfrac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i}^{n} (K_{h,ij}^{x})^2 (K_{h,ij}^{y})^2$。

在定理12.11中定义的是 $\sigma_f^2 = \kappa^{p+q} E[f_1(X_1)^2] E[f_2(Y_i)^2]$ 的一致估计量。

提示:注意在 H_0^f 假定条件下,X 和 Y 是独立的,

$$\begin{aligned} E[\hat{\sigma}_f^2] &= 2E[K_{h_x}(X_i - X_j)]E[K_{h_y}(Y_i - Y_j)] \\ &= 2\kappa^p \{E[f_1(X_i)^2] + o(1)\} \kappa^q \{E[f_2(Y_i)^2] + o(1)\} \\ &= \sigma_f^2 + o(1) \end{aligned}$$

运用引理A.16证明 $\mathrm{var}(\hat{\sigma}_h^2) = o(1)$。

第 13 章　非平滑检验

第 12 章中曾讨论了条件矩检验，它是基于条件均值函数的非参数核估计。但是，我们可以构造一个基于非条件矩估计的一致检验，而不是用非参数核估计方法来估计条件均值，这就导致了一类所谓的非平滑检验。 397

注意条件矩检验的原假定 H_0：几乎确定 $E(u_i \mid X_i)=0$，等价于对于所有的 $M(\cdot)\in\mathcal{M}$，$E[u_i M(X_i)]=0$，其中$\mathcal{M}$表示一类有界的 Borel 可测（或几乎处处连续）函数。遵循 Bierens(1982)，Bierens 和 Ploberger(1997)，Stinchcombe 和 White(1998)，以及 Stute(1997)，考虑下面的非条件矩检验

$$E[u_i\,\mathcal{H}(X_i,x)]=0,\text{几乎所有 } x\in\mathcal{S}\subset\mathbb{R}^q \tag{13.1}$$

其中 S 是 X_i 的支撑集。给定$\{\mathcal{H}(X_i,x):x\in\mathcal{S}\}$的线性空间(linear span)在$\mathcal{S}$上的有界且连续函数空间上是稠密的（参见附录的定义 A.36），对于所有的 $M(\cdot)\in M$，等式(13.1)等价于条件 $E[u_i M(X_i)]=0$。我们假定加权函数$\mathcal{H}(\cdot,\cdot)$在$\mathcal{S}\times\mathcal{S}$上是有界的。在非参数统计文献中的一个流行的对加权函数的选择是$\mathcal{H}(X_i,x)=\mathbf{1}(X_i<x)$，这里的 $\mathbf{1}(\cdot)$ 是普通示性函数，见 Stute(1997)。其他的加权函数选择包括指数函数$\mathcal{H}(X_i,x)=\exp(\iota X_i'x)$ $(\iota=\sqrt{-1})$ 和 logistic 函数$\mathcal{H}(X_i,x)=1/[1+\exp(c-X_i'x)$ $(c\neq 0)$；进一步的讨论见 Stinchcombe 和 White(1998)。通过把式(12.5)中给出的条件矩检验转化为式(13.1)中给出的非条件矩检验，当我们构造这个检验时避免了非参数估计的需要。因此，基于式(13.1)的检验减轻了维数诅咒的问题。 398

一个可行的检验可以通过用$\sqrt{n}$乘以 $E[u_i\,\mathcal{H}(X_i,x)]$的对应样本值来获得，如：

$$J_n(x)=\sqrt{n}\left[\frac{1}{n}\sum_{j=1}^{n}u_j\,\mathcal{H}(X_j,x)\right]=\frac{1}{\sqrt{n}}\sum_j u_j\,\mathcal{H}(X_j,x) \tag{13.2}$$

容易看到对一个固定的 x 值,由 Lindeberg-Levy 中心极限定理可知 $J_n(x)$ 收敛于一个零均值有限方差的正态随机变量。为了确定 H_0^a 是否成立,需要检验式(13.1)是否对所有的 x 成立。因此,为了获得一个基于式(13.2)的一致检验,需要对所有的 x 考虑 $J_n(x)$,一个这样做的方法是所谓的积分条件矩检验(一个 Cramer-von Mises 形式的检验),有如下形式

$$\mathrm{CM} = \int J_n(x)^2 \mathrm{d}F(x) = \int \left[n^{-1} \sum_i u_i \mathcal{H}(X_i, x) \right]^2 \mathrm{d}F(x) \qquad (13.3)$$

通过应用中心极限定理,可以证明 $J_n(x)$ 收敛于一个零均值高斯过程(如 $J_\infty(\cdot)$)。然后,通过连续映射定理可以证明 $\mathrm{CM} \to \int J_\infty(x)^2 \mathrm{d}F(x)$,它提供了一个一致非平滑检验的基础。在实际操作中,u_i 是不可观测的,我们可以用一些一致估计量来替代。下面我们讨论一些在文献中出现的基于式(13.2)和式(13.3)的一致模型设定检验。

13.1 对参数回归函数形式的检验

当检验一个参数回归函数的正确设定时,即检验 H_0^a:几乎处处 $E(Y_i|x) = m(x, \gamma)$,我们用 $\hat{u}_i = Y_i - m(X_i, \hat{\gamma}_0)$ 来替代 u_i,并且构造一个式(13.2)的可行的形式。这样,对于 H_0^a 的检验统计量是基于

$$\hat{J}_n^a(x) = \frac{1}{\sqrt{n}} \sum_i \mathcal{H}(X_i, x) \hat{u}_i \qquad (13.4)$$

399

为了得到 $\hat{J}_n^a(x)$ 的渐近分布,通常我们首先选择一个范数,其中弱收敛将被定义。普遍使用的范数包括 Skorohod 范数(见 Stute(1997))和一些 L_2 范数。使用一个 L_2 范数使得渐近分析比使用 Skorohod 范数简单得多,因此我们将使用 L_2 范数来建立 $\hat{J}_n(\cdot)$ 的弱收敛。在13.7节中我们证明了以下结论:

定理 13.1 在一些标准正则性条件下,$\hat{J}_n(\cdot)$ 弱收敛于一个零均值高斯过程 $J_\infty(\cdot)$,其具有13.7节中给出的协方差结构。

利用 $\hat{J}_n^a(x)$ 我们可以构造由下式给出的一个 Cramer-von Mises(CM)形式统计量来检验 H_0^a,

$$\mathrm{CM}_n^a = \int [\hat{J}_n^a(x)]^2 \mathrm{d}F_n(x) = \frac{1}{n} \sum_{i=1}^n [\hat{J}_n^a(X_i)]^2 \qquad (13.5)$$

其中 $F_n(\cdot)$ 是 $X_1, \cdots, X_n$ 的经验分布。给定 $\hat{J}_n^a(\cdot)$ 缓慢收敛于 $J_\infty^a(\cdot)$,应用连续映射定理(见习题13.2),得到 $\mathrm{CM}_n^a \xrightarrow{d} \int [J_\infty^a(x)]^2 F(\mathrm{d}x)$,其中 $F(\cdot)$ 是 X_i

的累积分布函数。

正如 Bierens 和 Ploberger(1997),Stute(1997)以及 Wang(2000)中所述,我们能够证明$\int [J_{\infty}^{a}(x)]^2 F(dx)$可以写成无限个加权(独立的)$\chi_1^2$ 随机变量的和,这里的权重取决于(X_i, Y_i)的未知分布。因此,列出 CM_n 的临界值是不可能的。但是基于残差的原始自举方法可以用来近似估计 CM_n^a 的有限零分布的临界值,原始自举法的误差 u_i^* 是通过一个两点分布生成的,即以概率$(1+\sqrt{5})/[2\sqrt{5}]$,$u_i^* = [(1-\sqrt{5})/2]\hat{u}_i$;以概率$(\sqrt{5}-1)/[2\sqrt{5}]$,$u_i^* = [(\sqrt{5}+1)/2]\hat{u}_i$。利用$\{u_i^*\}_{i=1}^n$,我们根据原模型 $Y_i^* = m(X_i, \hat{\gamma}) + u_i^*$ 生成 Y_i^*。利用自举样本$\{Y_i^*, X_i\}_{i=1}^n$,我们估计基于原模型的 γ,并且用 $\hat{\gamma}$ 表示得到的估计值。我们得到自举残差 $\hat{u}_i^* = Y_i^* - m(X_i, \hat{\gamma}^*)$,并且通过用 $\hat{u}_i^*$ 来替代 $\hat{J}_n^a(x)$ 中的 $\hat{u}_i$ 而得到 $\hat{J}_n^{a*}(x)$的自举统计量。使用 $\hat{J}_n^*(\cdot)$我们可以计算 CM_n^a 统计量的一个自举形式,即 $CM_n^{a*} = n^{-1}\sum_i [\hat{J}_n^{a*}(X_i)]^2$。条件于随机样本$\{(X_i, Y_i)\}_{i=1}^n$,在 H_o 假定下,CM_n^{a*} 依概率和依分布收敛于$\int [J_{\infty}^{a}(x)]^2$。因此,$CM_n^{a*}$ 可以用来近似 CM_n^a 的零分布。“依概率依分布收敛”的定义在式(12.8)中给出。 400

可以证明,CM_n^a 检验可以检测以 $O_p(n^{-1/2})$的速度逼近初始模型的 Pitman 局部替代。平滑和非平滑检验看上去是相当不同的,因为它们是由不同的原理得到的。但是,Fan 和 Li(2000)证明了这两种检验在实际上是非常相关的。特别是,他们证明了如果我们把在12章中定义的核检验统计量 I_n^a 中的 h_s 固定为一个常数,例如 $h_s = 1(s=1,\cdots,q)$,并且如果核函数 $w_s(\cdot)$满足一些附加的条件,那么,I_n^a 就变成一个非平滑检验,即

$$I_n^a(h=1) = \frac{1}{n(n-1)}\sum_i \sum_{j\neq i} \hat{u}_i \hat{u}_j K(X_i - X_j) \tag{13.6}$$

是一个非平滑检验,其中 $K(X_i - X_j) = \prod_{s=1}^{q} k(X_{is} - X_{js})$是对所有 $s=1,\cdots,q$ 满足 $h_s = 1$ 的核函数的乘积。Fan 和 Li 证明了对于常用的核函数,包括均匀、标准正态、三角和 Epanechnikov 核,$I_n^a(h=1)$检验确实是一个一致检验。因此,一个平滑和非平滑检验的唯一不同是前者使用减小的 h_s(当 $n\to\infty$ 时,$h_s\to 0$)。基于 Pitman 局部替代的分析说明了非平滑检验比平滑检验有效得多,因为前者可以检测以 $O_p(n^{-1/2})$的速度逼近初始模型的 Pitman 局部替代,而后者只能检测以 $O_p(((n(h_1\cdots h_q)^{1/2})^{-1/2})$的速度逼近初始模型的 Pitman 局部替代,这比 $O_p(n^{-1/2})$慢。但是,Fan 和 Li 证明在高频局部替代(Rosenblatt(1975))下,平滑

检验要比非平滑检验更有功效。Fan 和 Li 的蒙特卡洛模拟证明了确实是这种情况(也见 Andrews(1997))。因此,非平滑检验和平滑检验互补。

式(13.6)给出的非平滑检验的一个问题是它对于不同大小的数据不是不变的。[①] 为了修正这一问题,我们建议使用

$$I_n^a(h=1)=\frac{1}{n(n-1)}\sum_i\sum_{j\neq i}\hat{u}_i\hat{u}_jK\left(\frac{X_i-X_j}{x_{sd}}\right) \tag{13.7}$$

401

其中 $K((X_i-X_j)/x_{sd})=\prod_{s=1}^{q}k((X_{is}-X_{js})/x_{s,sd})$,其中 $x_{s,sd}$ 是 $\{X_{is}\}_{i=1}^n$ 的样本标准差。

13.2 概率密度函数的等价性检验

我们在 12.2 节中讨论了两个未知的概率密度函数的等价性检验。原假定是 H_0^b:几乎处处 $f(x)=g(x)$。Anderson 等(1994)建议使用一个固定的窗宽 h 来检验 H_0^b。不失一般性,我们在 I_n^b 中使用 $h=1$,得到一个非平滑检验统计量如下:

$$J_n^b=\frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}K_{ij}^x+\frac{1}{n_2^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}K_{ij}^y-\frac{2}{n_1n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}K_{ij}^{x,y} \tag{13.8}$$

其中 $K_{ij}^x=K(X_i-X_j)$,$K_{ij}^y=K(Y_i-Y_j)$,$K_{ij}^{x,y}=K(X_i-Y_j)$。在 n_1 和 $n_2\to\infty$ 和 n_1/n_2 偏离零并趋于无穷的假定下,Anderson 等人证明了 $T_{n,ns}^b=n_1\left[J_n^b-\left(\frac{1}{n_1}+\frac{1}{n_2}\right)k(0)^q\right]$ 收敛于一个独立 χ_1^2 分布的无穷项和。自举形式的 J_n^b 可以如 13.1 节(对所有的 $j=1,\cdots,q$,都有 $h_j=1$)中同样的形式产生。用 J_n^{b*} 表示得到的 J_n^b 的自举形式。自举统计量是 $T_{n,ns}^{*b}=n_1\left[J_n^{*b}-\left(\frac{1}{n_1}+\frac{1}{n_2}\right)k(0)^q\right]$,并且自举统计量的经验分布可以用来逼近 T_n^b 的零分布。并且,实际上,我们可能希望用 $K((X_i-X_j)/y_{sd})$ 来替代 $K_{ij}^x=K(X_i-X_j)$,用 $K((Y_i-Y_j)/y_{sd})$ 来替代 K_{ij}^y,用 $K(X_i/x_{sd}-Y_j/y_{sd})$ 来替代 K_{ij}^{xy},来使得对于不同的数据大小的检验都是始终如一的。

---

① 可直接使用,比如说 $h=n^{-1/5}$ 与 $h=x_{sd}n^{-1/5}$(其中 x_{sd} 是 $\{x_i\}_{i=1}^n$ 的样本标准差)相比,后者更受欢迎,因为它关于 x 的度量(即标准差)是不变的。

13.3 非参数显著性检验

接下来,原假定与第12章中讨论的 H_0^d 相同,即 H_0^d:几乎处处 $E(Y_i|w,z)=r(w)$,其中 $r(w)=E(Y_i|w)$。定义 $u_i=(Y_i-r(W_i))f(X_i)(x=(w,z))$。这样,$H_0^d$ 与 $E(u_i|X_i)=0$ 等价。这样,我们可以构造一个基于 $J_n^d(x)=n^{-1}\sum_i u_i \mathcal{H}(X_i,x)$ 的检验统计量。Chen 和 Fan(1999)使用了一个平滑加权函数 $\mathcal{H}(X_i,x)$,Delgado 和 Manteiga(2001)采用的是 $\mathcal{H}(X_i,x)=\mathbf{1}(X_i\leqslant x)$。为了得到一个可行的检验统计量,我们可以使用核估计量 $\hat{r}(X_i)=n^{-1}\sum_i Y_jK_{h_w,ij}/\hat{f}_w(W_i)$ 和 $\hat{f}_w(W_i)=n^{-1}\sum_i Y_jK_{h_w,ij}$ 来替代 $r(W_i)$ 和 $f_w(W_i)$,得到: 402

$$\hat{J}_n^d(x)=\frac{1}{n}\sum_{i=1}^n \hat{u}_i\hat{f}(X_i)\,\mathcal{H}(X_i,x) \tag{13.9}$$

其中 $\hat{u}_i\hat{f}(X_i)=(Y_i-\hat{r}(W_i))\hat{f}_w(W_i)$。Chen 和 Fan 使用了 L_2 范数,而 Delgado 和 Manteiga 使用的是 Skorohod 范数,可以证明无论是 L_2 范数还是 Skorohod 范数,$\hat{J}_n^d(\cdot)$ 在 H_0^d 下都收敛于一个零均值高斯过程。

注意 $\hat{J}_n^d(\cdot)$ 检验是一个不完全的非平滑检验,因为还需要平滑原模型来获得 u_i 和 $f_w(\cdot)$ 的一致估计。但是与第12章中考虑的平滑检验 $\hat{I}_n^d$ 不同,它不平滑对立模型。因此,它减轻了维数诅咒问题。一个 CM 形式的统计量由下式给出:

$$\mathrm{CM}_n^d=\int[\hat{J}_n^d(x)]^2F_n(\mathrm{d}x)=\frac{1}{n}\sum_i[\hat{J}_n^d(X_i)]^2 \tag{13.10}$$

CM_n^d 的自举形式由下式给出:

$$\mathrm{CM}_n^{d*}=\frac{1}{n}\sum_i[\hat{J}_n^{d*}(X_i)]^2 \tag{13.11}$$

其中 $\hat{J}_n^{d*}(X_i)=n^{-1}\sum_{j=1}^n \hat{u}_j^* H(X_j,X_i)$,$\hat{u}_j^*=Y_i^*-\hat{r}^*(W_i)$,$Y_j^*=\hat{r}^*(W_i)+u_i^*$,$\hat{r}^*(W_i)=n^{-1}\sum_i Y_j^*K_{h_w,ij}/\hat{f}_w(W_i)$,并且 u_i^* 是由 $\hat{u}_i=Y_i-\hat{m}(W_i)$ 产生的两点原始自举误差。Delgado 和 Manteiga(2001)证明了自举方法在中等样本大小的背景下表现很好。

13.4 条件累积分布函数的 Andrews 检验

Andrews(1997)提出了一个非平滑检验来检验一个参数条件累积分布函数

的正确设定。Andrews 把他的方法叫作条件 Kolmogorov(CK)检验。令 $f(y|x,\theta)$ 代表参数条件概率密度函数。参数条件累积分布函数由下式给出：

$$F(y|x,\theta)=\int \mathbf{1}(v<y)f(v|x,\theta)\mathrm{d}\mu(v) \tag{13.12}$$

403 其中 $\mathrm{d}\mu(v)=\mathrm{d}v$ 并且 Y 是一个连续变量，而当 Y 是离散变量时令 $\mathcal{Z}_y$ 表示 Y 的值域，则 $\int \mathrm{d}\mu(v)=\sum_{v\in\mathcal{D}_y}$。$Y$ 也可以由连续和离散值混合组成。令 $H(\cdot|\cdot)$ 表示对于给定 X_i 下 Y_i 的条件累积分布函数，令 $G(\cdot)$ 表示 X_i 的边际累积分布函数。原假定是

$$H_0^i:H(y|x)=F(y|x,\theta)\text{a. e. }(\text{in}(y,x))\text{ 某些 }\theta\in\Theta \tag{13.13}$$

其中 Θ 是 $\mathbb{R}^L$ 上的一个紧子集。令 $\hat{H}_n(z)$ 表示 $Z_i=(X_i,Y_i),i\leqslant n$ 的经验累积分布函数。即

$$\hat{H}_n(z)=\frac{1}{n}\sum_{i=1}^n \mathbf{1}(Z_i<z) \tag{13.14}$$

令 $\hat{G}_n(x)$ 表示 $X_i,i\leqslant n$ 的经验累积分布函数，即

$$\hat{G}_n(x)=\frac{1}{n}\sum_{i=1}^n \mathbf{1}(X_i<x) \tag{13.15}$$

令 $\hat{\theta}$ 表示一个基于原模型的 θ 的估计量，令 $\hat{F}_n(z,\hat{\theta})$ 表示基于参数条件累积分布函数 $F(\cdot|\cdot,\hat{\theta})$ 和经验累积分布函数 $\hat{G}_n(\cdot)$，$\{Z_i,i\leqslant n\}$ 的半参数/半经验累积分布函数表示如下：

$$\hat{F}_n(z,\hat{\theta})=\frac{1}{n}\sum_{i=1}^n F(y|X_i,\hat{\theta})\mathbf{1}(X_i\leqslant x) \tag{13.16}$$

Andrews(1997)提出了下面的 CK 检验统计量：

$$\begin{aligned}\mathrm{CK}_n&=\sqrt{n}\max_{1\leqslant j\leqslant n}|\hat{H}_n(Z_j)-\hat{F}_n(Z_j,\hat{\theta})|\\&=\max_{1\leqslant j\leqslant n}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^n[\mathbf{1}(Y_i\leqslant Y_j)-F(Y_j|X_i,\hat{\theta})]\mathbf{1}(X_i\leqslant X_j)\right|\end{aligned} \tag{13.17}$$

Andrews(1997)证明了 CK_n 的渐近分布是冗余参数依赖的，它依赖于 θ_0 和 $G(\cdot)$。结果，CK_n 的渐近临界值不能被列出。Andrews 建议使用下面的参数自举过程来获得 CK_n 的临界值：

404 (i) 对于 $i=1,\cdots,n$，根据参数条件概率密度函数 $f(Y_i|X_i,\hat{\theta})$(或 $F(Y_i|X_i,\hat{\theta})$)产生 Y_i^*。

(ii) 计算自举检验量 CK_n^*，这里的 CK_n^* 除了在 CK_n 中的 Y_i 被替代成了 Y_i^* 之外与 CK_n 是一样的。

(iii) 重复(i)和(ii)多次,比如 $B=399$ 次,并得到 $\{\mathrm{CK}_n^{*j}\}_{j=1}^{B}$ 的经验分布,它可以用来获得 CK_n 的临界值。

Andrews(1997)为上述的自举过程和模拟报告提供了辩护,他证明了参数自举方法在有限样本背景中表现良好。

13.5 序列相关的 Hong 检验

Hong(1999)提出了一类可以用于检验各种形式的序列相关的广义谱检验。令 $\{X_t\}$ 为一个严格平稳过程,有一个边际特征函数 $\phi(u)=E(\mathrm{e}^{iuX_t})$ 和配对联合特征函数 $\phi_j(u,v)=E[\mathrm{e}^{i(uX_t+vX_{t-|j|})}]$,其中 $i=\sqrt{-1}$,$u,v\in(-\infty,\infty)$,且 $j=0,\pm1,\cdots$定义转换变量 e^{iuX_t}和 $\mathrm{e}^{ivX_{t-|j|}}$之间的协方差函数为:

$$\sigma_j(u,v)\equiv\operatorname{cov}(\mathrm{e}^{iuX_t},\mathrm{e}^{ivX_{t-|j|}}) \tag{13.18}$$

可以直接证明,$\sigma_j(u,v)=\phi_j(u,v)-\phi(u)\phi(v)$。因为当且仅当 X_t 和 $X_{t-|j|}$ 独立时,对所有的 u、v 都有 $\phi_j(u,v)=\phi(u)\phi(v)$,$\sigma_j(u,v)$ 可以捕捉到 X_t 和 $X_{t-|j|}$之间任何类型的(序列)相关性。假定 $\sup_{u,v\in(-\infty,\infty)}\sum_{j=-\infty}^{\infty}|\sigma_j(u,v)|<\infty$,则 $\sigma_j(u,v)$的傅里叶变换存在,即

$$f(w,u,v)\equiv\frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\sigma_j(u,v)\mathrm{e}^{-ijw},\quad w\in[-\pi,\pi] \tag{13.19}$$

$f(w,u,v)$包含了在 $\{X_t\}$ 的所有滞后项中所有成对的序列相关的信息。因此,可以构造基于$f(\cdot)$的用于检验各种形式的序列相关的检验统计量。

Hong(1999,定理 1)证明了 $f(w,u,v)$ 可以通过下式被一致估计:

$$\hat{f}(w,u,v)=\frac{1}{2\pi}\sum_{j=1-n}^{n-1}k(j/p)\hat{\sigma}_j(u,v)\mathrm{e}^{-ijw} \tag{13.20}$$

其中 $\hat{\sigma}_j(u,v)=\hat{\phi}_j(u,v)-\hat{\phi}(u,0)\hat{\phi}(0,v)$,且 $\hat{\phi}_j(u,v)=(n-|j|)^{-1}\sum_{t=|j|+1}^{n}\mathrm{e}^{i(uX_t+vX_{t-|j|})}$ 是经验(成对的)特征函数,$p=p_n$ 是窗宽或滞后阶数,并且 $k(\cdot)$是一个核函数。p(即最小化$\hat{f}(\cdot)$的积分均方误差)的最优选择取决于$f(\cdot)$的平滑度和 k 的选择。对于 Bartlett 核,$p\sim n^{1/3}$;对于 Daniell 和 Parzen 核,$p\sim n^{2/5}$。Hong 也提出了一个选择最优 p 值的数据驱动方法。注意 Daniell 核是由 $k(z)=\sin(\pi z)/\pi z$,$z\in(-\infty,\infty)$给出的。

如果$\{X_t\}$是一个独立同分布序列,则对于所有的$|j|\geqslant1$,有 $\sigma_j(u,v)=0$。所以,$f(w,u,v)$是一个在 w 上的常数函数:

$$f_{(0)}(w,u,v)=\frac{1}{2\pi}\sigma_0(u,v),\quad w\in[-\pi,\pi]$$

这样，可以通过比较式(13.20)中定义的$\hat{f}(w,u,v)$和$\hat{f}_{(0)}(w,u,v)=\frac{1}{2\pi}\hat{\sigma}_0(u,v)$（见下面定义的$M(0,0)$检验）来检验序列相关。如果独立同分布假定被拒绝，可以通过比较如下式给出的$f(w,u,v)$和$\hat{f}_{(0)}(w,u,v)$的导数估计量进一步检验序列相关的其他特定形式。

$$\hat{f}_n^{(0,m,l)}(w,u,v)=\frac{1}{2\pi}\sum_{j=1-n}^{n-1}\left(1-\frac{|j|}{n}\right)^{1/2}k(j/p)\hat{\sigma}_j^{(m,l)}(u,v)\mathrm{e}^{-ijw} \quad (13.21)$$

$$\hat{f}_{(0)}^{0,m,l}(w,u,v)=\frac{1}{2\pi}\hat{\sigma}_0^{(m,l)}(u,v) \quad (13.22)$$

其中$\hat{\sigma}_j^{(m,l)}(u,v)=\partial^{m+l}\hat{\sigma}_j(u,v)/\partial^m u\partial^l v$，所有$m,l\geqslant 0$。

Hong(1999)提出了一类基于由下式给出的$\hat{f}_n^{(0,m,l)}(w,u,v)$和$\hat{f}_{(0)}^{(o,m,l)}(w,u,v)$之间的加权积分平方差的检验：

$$\iint_{-\pi}^{\pi}|\hat{f}_n^{(0,m,l)}(w,u,v)-\hat{f}_{(0)}^{(m,l)}(w,u,v)|^2\mathrm{d}w\mathrm{d}W_1(u)\mathrm{d}W_2(v)$$

$$=\frac{2}{\pi}\int\sum_{j=1}^{n-1}k^2(j/p)\left(1-\frac{j}{n}\right)|\hat{\sigma}^{(m,l)}(u,v)|^2\mathrm{d}W_1(u)\mathrm{d}W_2(v) \quad (13.23)$$

这里的等式遵循 Parseval 等式，其中$W_1(\cdot)$和$W_2(\cdot)$是下面要讨论的加权函数。

406 检验统计量是上面的二次形式的标准化形式：

$$M(m,l)=\left[\int\sum_{j-1}^{n-1}k^2(j/p)(n-j)|\hat{\sigma}_j^{(m,l)}(u,v)|^2\mathrm{d}W_1(u)\mathrm{d}W_2(v)-\hat{C}_0^{(m,l)}\sum_{j-1}^{n-1}k^2(j/p)\right] \quad (13.24)$$

其中$W_1(\cdot)$和$W_2(\cdot)$是被选择的加权函数，

$$\hat{C}_0^{(m,l)}=\int\hat{\sigma}^{(m,m)}(u,-u)\mathrm{d}W_1(u)\int\hat{\sigma}^{(l,l)}(v,-v)\mathrm{d}W_2(v)$$

$$\hat{D}_0^{(m,l)}=2\times\int|\hat{\sigma}_0^{(m,m)}(u,u')|^2\mathrm{d}W_1(u)\mathrm{d}W_2(u')\times\int|\hat{\sigma}_0^{(m,m)}(v,v')|^2\mathrm{d}W_1(v)\mathrm{d}W_2(v')$$

对于给定的(m,l)，$M(m,l)$在序列无关的原假定下渐近(单边)$N(0,1)$。

常用的假定包括选择$(m,l)=(0,0)$来检验序列相关的任意形式。如果拒绝一般序列无关的原假定，人们可以继续检验序列相关的特定形式。例如，选择$(m,l)=(1,0)$检验原假定$E(X_t|X_{t-j})=E(X_t)$，所有$j>0$，这被称为鞅差分(martingale differences)的检验。类似地，对于$l=1,2,3,4$的$(m,l)=(1,l)$检验

是否对于所有的 $j>0$ 都有 $\mathrm{cov}(Y_t, Y_{t-j}^l)=0$，它分别检验均值的线性性、均值的 ARCH 效应、均值偏度和均值峰度效应。下面我们提供了更多关于上面式(13.24)给出的对于 $(m,l)=(0,0)$ 和 $(m,l)=(1,0)$ 的一般 $M(m,l)$ 统计量的应用细节。

由于 $(m,l)=(0,0)$，Hong(1999)建议选择 $W_1(\cdot)=W_2(\cdot)=W_0(\cdot)$，其中 $W_0(\cdot)$ 是标准正态随机变量的累积分布函数。

那么，可以证明： 407

$$M(0,0)=\left\{\iint\left[\sum_{j=1}^{n-1}k^2(j/p)(n-j)\mid\hat{\sigma}_j(u,v)\mid^2\right]\mathrm{d}W_0(u)\mathrm{d}W_0(v)-\hat{C}^{(0,0)}\sum_{j=1}^{n-1}k^2(j/p)\right\}/\hat{\sigma}_{(0,0)} \tag{13.25}$$

其中 $\hat{C}^{(0,0)}=\left[\int\hat{\sigma}_0(u,-u)\mathrm{d}W_0(u)\right]^2$ 且

$$\hat{\sigma}_{(0,0)}^2=2\left[\iint\mid\hat{\sigma}_0(u,v)\mid^2\mathrm{d}W_0(u)\mathrm{d}W_0(v)\right]^2\sum_{j=1}^{n-2}k^4(j/p)$$

为了检验 $(m,l)=(1,0)$ 的鞅差分，Hong(1999)建议使用 $W_1(\cdot)=\delta(\cdot)$，$\delta(\cdot)$ 为 Dirac δ 函数(见附录中的定义 A.15)，其中 $W_2(\cdot)=W_0(\cdot)$。则

$$M(1,0)=\left\{\int\left[\sum_{j=1}^{n-1}k^2(j/p)(n-j)\left|\hat{\sigma}_j^{(1,0)}(0,v)\right|^2\right]\mathrm{d}W_0(v)-\hat{C}^{(1,0)}\sum_{j=1}^{n-1}k^2(j/p)\right\}\Big/\hat{\sigma}_{(1,0)} \tag{13.26}$$

其中

$$\hat{C}_0^{(1,0)}=\hat{R}_{1,0}\int\hat{\sigma}_0(u,-u)\mathrm{d}W_0(u)$$

$$\hat{\sigma}_{(1,0)}=\left[2\hat{R}_{1,0}^2\int\left|\hat{\sigma}_0^2(u,u)\right|^2\mathrm{d}W_0(u)\sum_{j=1}^{n-2}k^4(j/p)\right]^{1/2}$$

$$\hat{\sigma}_j^{(1,0)}=(n-j)^{-1}\sum_{t=j+1}^{n}X_t\left[\mathrm{e}^{ivX_{t-j}}-\hat{\phi}_j(0,v)\right]$$

$$\hat{R}_{1,j}=(n-j)^{-1}\sum_{t=j+1}^{n}[X_t-\overline{X}_t(j)][X_{t-j}-\overline{X}_2(j)]$$

$$\overline{X}_1(j)=(n-j)^{-1}\sum_{t=j+1}^{n}X_t$$

$$\overline{X}_2(j)=(n-j)^{-1}\sum_{t=j+1}^{n}X_{t-j}$$

Hong 和 Lee(2003)推荐使用自举法来更好地逼近检验统计量 $M(m,l)$ 的 408

有限样本零分布。令 $\hat{M}_{m,l}$ 为基于原始样本 $\{X_t\}_{t=1}^{n}$ 的统计量，$\hat{M}_{m,l}^{b}$ 为使用自举法样本 $\{X_t^b\}_{t=1}^{n}$ 的统计量。$\hat{M}_{m,l}$ 的自举法 P 值可由 $p_B = B^{-1}\sum_{b=1}^{B}\mathbf{1}(\hat{M}_{m,l}^{b} \geqslant \hat{M}_{m,l})$ 近似得到，其中 B 为自举法重复次数且 $\mathbf{1}(\cdot)$ 为示性函数。检验独立同分布的原假定（$(m,l)=(0,0)$），可使用朴素（重新取样）自举法；对于鞅差分的原假定可使用原始自举法检验，原始自举法以概率 r 取 $Y_t^b = aY_t$，以概率 $1-r$ 取 $Y_t^b = (1-a)Y_t$，而 $a=(1+\sqrt{5})/2$，$r=a/\sqrt{5}$。Hong 和 Lee 证明了当样本大小为 $n=100$ 时，自举法表现良好。

Hong 和 Lee(2003)采用了 Hong(1999)的对于不同 (m,l) 的值的广义谱检验；在不同的 (m,l) 值的情况下对 $W_1(\cdot)$ 和 $W_2(\cdot)$ 的相应选择见 Hong 和 Li 的表 1。Hong 和 Lee 对加拿大、德国、美国、日本和法国的日汇率数据进行了检验。一般发现是五个汇率表现出较强的序列相关性（拒绝原假定情形 $(m,l)=(0,0)$），而这些汇率的变化通常是序列不相关的（即无法拒绝原假定 $(m,l)=(1,1)$），但是它们明显不是鞅差分过程（拒绝原假定 $(m,l)=(1,0)$）。他们的结论表明可以预测汇率变化的平均值。他们进一步使用不同的非线性时间序列模型来预测汇率的变化，并且发现将几个不同预测模型组合在一起的方法可以得到最好的预测。

13.6 更多的非平滑检验

原则上，所有一致模型设定检验都可以使用非平滑方法来构造。例如，Delgado(1993)提出了对两个非参数回归曲线是否相同的非平滑检验，Bai 和 Ng(2001)考虑了一个关于条件对称性的非平滑检验，Fan(1994)提出了一个参数概率密度函数的非平滑检验，Fan 和 Gencay(1995)提出了一个线性回归模型对称性的一致性检验。

409

13.7 证 明

13.7.1 定理 13.1 的证明

我们首先给出一个引理。

引理 13.1 令 $Z_1(\cdot),\cdots,Z_n(\cdot)$ 为 L_2 空间（一个 Hilbert 空间）里的独立同分布零均值随机元素，有范数 $\|Z_i(\cdot)\|_\nu^2 = \int |Z_i(x)|^2 d\nu(x)$，其中在 X_i 的

支撑集$\mathcal{S}$上，$\nu(\cdot)$是一个有限测度。如果$E[|Z_i(\cdot)|_\nu^2]<\infty$，则$n^{-1/2}\sum_{i=1}^n Z_i(\cdot)$弱收敛于零均值高斯过程，其协方差函数由$\Omega(x,x')=E[Z_i(x)Z_i(x')]$给出。

证明：见 Politis 和 Romano(1994)的定理2.1，或者参见 van der Vaart 和 Wellner(1996,ex. 1.8.5, p.50)。注意$E[\|Z_i(\cdot)\|_\nu^2]<\infty$是保证过程$n^{-1/2}\sum_{i=1}^n Z_i(\cdot)$紧致的一个充分条件。

为了证明定理13.1，我们首先有式子：

$$n^{-1/2}\sum_i Z_i(\cdot)=n^{-1/2}\sum_i \hat{u}_i\,\mathcal{H}(X_i,\cdot)$$

其中$\hat{u}_i=Y_i-m(X_i,\hat{\gamma})$。假定$\hat{\gamma}$是$\gamma$的非线性最小二乘估计值。令$m_i^{(1)}=[\partial m(X_i,\gamma)/\partial\gamma]|_\gamma=\gamma_0$，且由$\hat{\gamma}-\gamma_0=O_p(n^{-1/2})$的事实和使用Talor级数展开，我们有$\hat{u}_i=Y_i-m(X_i,\gamma_0)-m_i^{(1)}(\hat{\gamma}-\gamma)+O_p(n^{-1})$。同时，根据非线性最小二乘理论我们知道

$$\begin{aligned}\hat{\gamma}-\gamma_0&=\left[n^{-1}\sum_i m_i^{(1)'}m_i^{(1)}\right]^{-1}\sum_i m_i^{(1)}u_i+O_p(n^{-1})\\&=\left\{E\left[m_i^{(1)'}m_i^{(1)}\right]\right\}^{-1}\sum_i m_i^{(1)}u_i+O_p(n^{-1})\end{aligned}\tag{13.27}$$

令$B=E\left[m_i^{(1)'}m_i^{(1)}\right]$，$\Phi(X_i,\cdot)=E\left[m_i^{(1)}\mathcal{H}(X_i,\cdot)\right]Bm_i^{(1)}$，我们有 410

$$\begin{aligned}\hat{J}_n^a(\cdot)&=n^{-1/2}\sum_i\hat{u}_i\,\mathcal{H}(X_{i,\cdot})\\&=n^{-1/2}\sum_i\hat{u}_i\,\mathcal{H}(X_{i,\cdot})+\left[n^{-1}\sum_i m_i^{(1)}\mathcal{H}(X_{i,\cdot})\right]B\left[n^{-1/2}\sum_i m_i^{(1)}u_i\right]+O_p(n^{-1})\\&=n^{-1/2}\sum_i\hat{u}_i\,\mathcal{H}(X_{i,\cdot})+E\left[m_i^{(1)}\mathcal{H}(X_{i,\cdot})\right]B\left[n^{-1/2}\sum_i m_i^{(1)}u_i\right]+O_p(n^{-1/2})\\&=n^{-1/2}\sum_i[\mathcal{H}(X_{i,\cdot})+\phi(X_{i,\cdot})]u_i\\&=n^{-1/2}\sum_i Z_i(\cdot)+O_p(n^{-1/2})\end{aligned}$$

其中$Z_i(\cdot)=[\mathcal{H}(X_i,\cdot)+\phi(X_i,\cdot)]u_i$。容易看出$E[\|Z_i(\cdot)\|_\nu^2]<\infty$(见习题13.1)。因此，由引理13.1可知$\hat{J}_n^a(\cdot)$收敛于一个零均值高斯过程，其协方差结构由$\Omega(x,x')=E[Z_i(x)Z_i(x')]$给出，其中$Z_i(x)=[\mathcal{H}(X_i,x)+\phi(X_i,\cdot)]u_i$。

□

13.8 习 题

习题 13.1 令 $Z_i(\cdot)=[\mathcal{H}(X_i,\cdot)+\phi(X_i,\cdot)]u_i$，证明 $E[\|Z_i(\cdot)\|_\nu^2]<\infty$。

习题 13.2 由定理 13.1，我们知道 $\hat{J}_n^a(\cdot)$ 收敛于 $J_\infty^a(\cdot)$，并且由连续映射定理（见附录 A），我们知道：

$$\int\left[\hat{J}_n^a(x)\right]^2\mathrm{d}\nu(x)\to\int\left[\hat{J}_\infty^a(x)\right]^2\mathrm{d}\nu(x) \tag{13.28}$$

在 $F(\cdot)$ 关于 ν 绝对连续的假定下，式(13.28)意味着（以 $F(\cdot)$ 取代 ν）：

$$\int\left[\hat{J}_n^a(x)\right]^2\mathrm{d}F(x)\to\int[J_\infty^a(x)]^2\mathrm{d}F(x) \tag{13.29}$$

411 证明（$\mathrm{d}F_n(\cdot)$ 对 $\mathrm{d}F(\cdot)$）：

$$\int\left[\hat{J}_n^a(x)\right]^2\mathrm{d}F_n(x)-\int[\hat{J}_n^a(x)]^2\mathrm{d}F(x)=o_p(1) \tag{13.30}$$

注意式(13.29)和式(13.30)意味着

$$\begin{aligned}\mathrm{CM}^a&\equiv\int\left[\hat{J}_n^a(x)\right]^2\mathrm{d}F_n(x)\\&=\int\left[\hat{J}_n^a(x)\right]^2\mathrm{d}F(x)+o_p(1)\\&\to\int\left[J_\infty^a(x)\right]^2\mathrm{d}F(x)\ \text{依分布收敛}\end{aligned}$$

第4部分　非参数近邻和序列方法

第14章 k 近邻估计方法

我们将在第15章研究非参数序列估计,它构成了本书的核心部分非参数核方法的一个替代。在这一章中,我们考虑另一个选择,称为近邻估计,它也被广泛使用,是已有的非参数方法的一种。 415

使用近邻估计的主要好处在于它能够自动适应可用的局部信息的数量。也就是说,局部信息量越大,平滑发生的区域越小。许多现有的文献表明,不变窗宽估计并不具有这个特点。但是,k 近邻估计方法(k-nearest neighbor method,以下简称“k-nn”)也有一些相对于核方法的缺点。例如,k-nn 方法不具有完全平滑掉非相关变量的能力。因此,k-nn 方法应该被看成是我们在前面的章节看到的流行的核估计方法的一个补充。

14.1 密度估计:一元情形

在式(1.6)中定义的均匀核估计量使用落在区间 $[x-h,x+h]$ 内的观测值来估计 $f(x)$。这个估计量的特点之一是采用固定平滑参数 h,也就是说 h 是一个常数且与 x 不相关。对于给定的点 x,当 $f(x)$ 较大时,则比 $f(x)$ 较小时有更多的数据落在区间 $[x-h,x+h]$ 内。比如说,在分布的尾部有相当少的数据落在区间 $[x-h,x+h]$ 内。但是,不变窗宽的使用可能导致在数据支撑集的一些范围内过度平滑而在其他地方平滑不足,比如在分布的尾部。 416

一个可供选择的非参数估计量,称为近邻密度估计量,可以用来规避这个问题。不是使用一个不变的窗宽,其中落在区间 $[x-h,x+h]$ 内的可观察数据数目会变化,一个简单的近邻方法使用位于 x 附近的 k 个观测值来估计 $f(x)$,其中 k 是一个固定的整数。也就是说,这个简单的估计量使用一个随 x 变动的窗宽,而不是保

持用于构造密度估计的观测数目不变。这种方法的优点之一是其局部适应性。也就是说，如果$f(x)$值小，间隔将会变大，反之亦然。不变窗宽核方法没有具备这样的性质。

核估计量和k-nn估计量的一个不同点是公式(14.12)定义的窗宽R_x现在是随机的。因此，k-nn估计量的渐近分析要比不变窗宽核估计量的渐近分析复杂。

设$X_1,X_2,\cdots,X_n$为独立同分布的随机观测值，它们的有界连续密度函数为$f(x)$。k-nn估计量最简单的形式是使用k个(离x)最邻近的观测值来估计。当x为q维向量时，我们将使用欧几里得距离来衡量数据点之间的距离。

定义14.1 令$R_x=R_n(x)$表示在x_i中的x和x的第k个近邻点之间的欧几里得距离。

因为$R_x=R_n(x)$依赖于n，为了记号的简洁，我们将隐藏记号n。对x的一个固定值，一个简单的k-nn估计量可以给出为：

$$\tilde{f}(x)=\frac{1}{nR_x}\sum_{i=1}^{n}\left(\frac{1}{2}\right)\mathbf{1}\left(\frac{|x-X_i|}{R_x}\leqslant 1\right)=\frac{k}{2nR(x)} \tag{14.1}$$

其中$\mathbf{1}(\cdot)$为通常的示性函数，其保证了只有最邻近x的k个观测值可用于估计$f(x)$。

为了得到$f(x)$的一致估计量，我们需要选择$k=k(n)$，使得当$n\to\infty$时，有$k\to\infty$且$k/n\to 0$。这里，k/n起的作用类似于核估计量的固定平滑参数h。条件$k\to\infty$和$k/n\to 0$对应于核估计里的$nh\to\infty$和$h\to 0$。

417

对于多元估计，$x\in\mathbb{R}^q$，我们将使用$\|\cdot\|$定义欧几里得范数，即

$$\|x-X_i\|=\sqrt{(x_1-X_{1i})^2+\cdots+(x_q-X_{qi})^2}$$

因此，一个简单的k-nn密度估计量由下式给出：

$$\tilde{f}(x)=\frac{1}{nR_x^q}\sum_{i=1}^{n}\frac{1}{c_0}\mathbf{1}\left(\frac{\|x-X_i\|}{R_x}\leqslant 1\right)=\frac{k}{c_0nR_x^q} \tag{14.2}$$

其中$c_0=\pi^{q/2}\Big/\left[\Gamma\left(\frac{q+2}{2}\right)\right]$是$\mathbb{R}^q$里单位球的体积。$\Gamma(\cdot)$是由$\Gamma(\alpha)=\int_0^\infty t^{\alpha-1}e^{-t}\mathrm{d}t$定义的$\Gamma$函数。通过$\Gamma(z+1)=z\Gamma(z)$，$\Gamma(1/2)=\sqrt{\pi}$和$\Gamma(1)=1$，对任何正整数值$q$，可以很容易地算出$c_0$。

当加权函数在球体边界不连续时，在式(14.2)中定义的$\tilde{f}(x)$赋予落在球体$\|X_i-x\|\leqslant R(x)$内的观测值相同的权重$(1/c_0)$，而球体外的观测值权重为0。连续的加权函数也可以使用，例如标准正态概率密度函数，即$w(v)=(2\pi)^{-q/2}\mathrm{e}^{-\|v\|^2/2}$，使得当$\|v\|$增加时，加权函数$w(v)$单调递减。使用一个一般的加权函数$w(\cdot)$，可以通过下式来估计$f(x)$：

$$\hat{f}(x) = \frac{1}{nR_x^q}\sum_{i=1}^{n} w\left(\frac{\| x - X_i \|}{R_x}\right) \tag{14.3}$$

其中 $w(\cdot)$是有界、对称、非负的积分加权函数，

$$\int_{R^q} w(v)\,\mathrm{d}v = 1$$

当 $w(\cdot)$为均匀概率密度函数(即如果 $\|v\| \leqslant 1, w(v) = 1/c_0$；如果 $\|v\| > 1, w(v) = 0$)时，在式(14.3)中给出的$\hat{f}(x)$退化为式(14.2)中定义的$\tilde{f}(x)$，为了简化下面的推导，我们施加约束，$\|v\| \geqslant 1$ 时，$w(v) = 0$。

假定 14.1　$w(\cdot)$是一个有界、对称、非负函数，对于 $\|v\| \geqslant 1$ 有 $w(v) = 0$，$\int w(v)\,\mathrm{d}v = 1$，$\int w(v) vv' dv = \kappa_2 I_q$，$\int w^2(v)\,dv = \kappa$，$\int w^2(v) vv'\mathrm{d}v = \kappa_{22} I_q$，$I_q$ 是 q 维单位矩阵，κ_2、κ_{22}都是有限的正常数。

接下来，我们对$f(x)$和 k 施加一些温和的条件。　418

假定 14.2

(i) $f(x)$二阶可微。

(ii) 当 $n \to \infty$ 时，$k \to \infty$ 且 $k/n \to 0$。

定理 14.1　在假定 14.1 和 14.2 条件下，我们有

(i) $\mathrm{bias}(\hat{f}(x)) = \dfrac{c_1 \mathrm{tr}[f^{(2)}(x)]}{f(x)^{2/q}}\left(\dfrac{k}{n}\right)^{2/q} + O((k/n)^{3/q})$

(ii) $\mathrm{var}(\hat{f}(x)) = \dfrac{c_2 f(x)^2}{k} + o(1/k)$

其中 $c_1 = (1/2)\,\kappa_2 c_0^{-2/q}$，$\kappa_2 = \int w(v)^2 v_s^2 \mathrm{d}v$，$v_s$ 为 v 的第 s 个分量，$c_2 = c_0\left[\int w^2(v)\,\mathrm{d}v\right]$。

对定理 14.1 的证明在 14.9 节给出。

对比定理 14.1 与式(1.13)、式(1.14)的结果，可以观察到$(k/n)^{1/q}$相当于$f(x)$的核估计量中使用的窗宽 h(或者$(k/n) \sim h^q$)。很容易看到平衡平方误差和方差的 k 的最优速度为 $O(n^{4/(4+q)})$阶。如果我们采用 k 的这个最优阶，则由定理 14.1 我们立即有

$$\begin{aligned}\mathrm{MSE}(\hat{f}(x)) &= \frac{[c_1 \mathrm{tr}(f^{(2)}(x))]^2}{[f(x)]^{4/q}}\left(\frac{k}{n}\right)^{4/q} + \frac{c_2 f^2(x)}{k} + (\mathrm{s.o.}) \\ &= O((k/n)^{4/q} + (1/k))\end{aligned} \tag{14.4}$$

对于第 1 章概述的核密度估计，误差项为 $O(h^2)$，它不依赖于 q 及 x 的维度。而方差 $O((nh^q)^{-1})$依赖于 q，所以导致了核估计的维数诅咒。但是，k-nn

方法的结论是不同的，因为方差项为 $O(1/k)$，其不依赖于 q，而误差项为 $O((k/n)^{2/q})$，也不依赖于 q，正是 q 引起了维数诅咒。

可以看到在偏误项分母中包含了 $f(x)^{2/q}$，因此，在分布的尾部 $f(x)$ 的值较小，平方偏误将是均方误差的重要成分。

Mack 和 Rosenblatt(1979)建立了以下关于 $\hat{f}(x)$ 的渐近正态结论。

419

定理 14.2 除了假定 14.1 之外，还假定 $f(x)$ 三阶可微，并且当 $n\to\infty$ 时，$k^{1/2}(k/n)^{3/q}\to 0$，则我们有

$$k^{1/2}\left(\hat{f}(x)-f(x)-(k/n)^{2/q}\frac{c_1\mathrm{tr}[f^{(2)}(x)]}{[f(x)]^{2/q}}\right)\xrightarrow{d}N(0,c_2f^2(x)) \quad (14.5)$$

核密度估计和 k-nn 密度估计之间关系的一个优秀的综述可以参见 Mack 和 Rosenblatt(1979, pp. 13—14)。

k-nn 方法的一个问题是它们对称地处理所有的成分。为了理解这一点，考虑 $q=2$ 的情况。如果 X_1 的数据范围(支撑集)远远大于 X_2，则 $\|X_i-x\|\leqslant R_x$ 可能包含了大部分有相同的 X_{i1} 和 x_1 但 X_{i2} 和 x_2 不同的点，这是不理想的。在实践中，我们可以先对数据进行标准化。使每一个成分都具有样本均值零和样本单位方差(例如，$(X_{is}-\bar{x}_s)/sd(X_{is})$)，或者对每个变量进行标准化从而使它们的取值在 0 和 1 之间，例如，$(X_{is}-\min_i X_{is})/(\max_i X_{is}-\min_i X_{is})$。后一种转变方法根据数据成分的范围使数据相对称，与基于没有转变的数据相比，将导致更好的有限样本的表现。但是，这仍然没有解决 x_s 潜在地对 $f(x)$ 有不同影响的问题。比方说，$f(x)$ 对于 x_1 是相对平稳的，而随着 x_2 的变化而变动剧烈。理想地，对于每一个变量 x_s，应该采用不同的 k_s，我们将在 14.8 节讨论这种可能性。

14.2 回归函数的估计

给定下面的非参数回归模型：

$$Y_i = g(X_i) + u_i \quad (14.6)$$

我们可以通过下式使用 k-nn 方法来估计回归函数 $g(x)=E(Y|x)$：

$$\hat{g}(x)=\frac{(nR_x^q)^{-1}\sum_{i=1}^{n}Y_{iw}\left(\frac{x-X_i}{R_x}\right)}{(nR_x^q)^{-1}\sum_{i=1}^{n}w\left(\frac{x-X_i}{R_x}\right)} \quad (14.7)$$

注意由于当加权函数乘以一个正的常数不改变其回归估计值时，假定式(14.1)中的 $\int w(v)\,\mathrm{d}v=1$ 可以由 $\int w(v)\,\mathrm{d}v=a>0$ 取代。

420

接下来的定理给出了 $\hat{g}(x)$ 的偏误首项和方差首项。

定理 14.3　除了假定 14.1 和 14.2 之外，还假定 $g(x)$ 二阶可微，那么我们有

$$\text{bias}(\hat{g}(x)) = c_3\left(\frac{k}{n}\right)^{2/q}\frac{2g^{(1)}(x)'f^{(1)}(x)+f(x)\text{tr}[g^{(2)}(x)]}{f(x)^{(2+q)/q}} + o(k^{-1/2}) + o((k/n)^{2/q})$$

$$\text{var}(\hat{g}(x)) = \frac{c_4\sigma^2(x)}{k} + o(k^{-1})$$

其中 $c_3 = \kappa_2/[2c_0^{2/q}]$，$\kappa_2 = \int w(v)v_s^2\text{d}v$，$c_4 = c_0[\int w^2(v)\text{d}v]$。

证明过程参见 Mack(1981)。

对比定理 14.3 和定理 2.2，我们再一次看到 $(k/n)^{1/q}$ 对应于 $g(x)$ 核估计量的 h 值。

定理 14.3 意味着均方误差收敛速度为

$$\text{MSE}(\hat{g}(x)) = [\text{bias}(\hat{g}(x))]^2 + \text{var}(\hat{g}(x)) = O\left(\left(\frac{k}{n}\right)^{4/q} + k^{-1}\right) \tag{14.8}$$

Mack(1981)也建立了以下的渐近正态结果：

定理 14.4　除了假定 14.1 和 14.2 之外，还假定 $g(x)$ 三阶可微，当 $n\to\infty$ 时，$k^{1/2}(k/n)^{3/q}\to 0$，则我们有

$$k^{1/2}(\hat{g}(x) - g(x) - (k/n)^{2/q}B_g(x)) \xrightarrow{d} N(0, c_4\sigma^2(x)) \tag{14.9}$$

其中 $B_g(x) = c_3\{2g^{(1)}(x)'f^{(1)}(x) + g(x)\text{tr}[f^{(2)}(x)]\}/[f(x)^{(2+q)/q}]$。

如之前提到的，对于多维变量 X 不能直接对原始数据采用 k-nn 估计。这是因为 x 的不同成分的值域一般是不同的。一个普遍的做法是先对 X_{is} 进行标准化，比方说，有零均值和单位标准差使得所有 X_{is} 具有可比较的值域。接下来对标准化了的变量应用欧几里得距离得到 k-nn 平滑参数。

14.3　局部线性 k-nn 估计

421

在这一节我们考虑这样的问题，使用局部线性 k-nn 方法来估计一个未知回归函数。考虑非参数回归模型

$$Y_i = g(X_i) + u_i, \quad i = 1,\cdots,n \tag{14.10}$$

其中 $X_i \in \mathbb{R}^q$，我们感兴趣的是估计未知函数 $g(x)$ 和它的导数，$\beta(x) \stackrel{\text{def}}{=} \partial g(x)/\partial(x) \equiv g^{(1)}(x)$。令 $\delta(x) = (g(x), (\beta(x))')'$。使用 Talor 级数对 $g(X_i)$ 关于 x 进行展开，我们得到 $g(X_i) = g(x) + (X_i - x)'\beta(x) + T(X_i, x)$，其中 $T(X_i, x) =$

$g(X_i)-g(x)-(X_i-x)'\beta(x)$。因此,我们可以记

$$Y_i = (1,(X_i-x)')\delta(x) + T(X_i,x) + u_i \tag{14.11}$$

令 $w(\cdot)$为有界、对称加权函数,且当 $\|v\| \geq 1$ 时,有 $\int w(v)dv=1, w(v)=0$。对 $\delta(x)$的局部线性 k-nn 估计可以由 $\hat{\delta}(x)$给出:

$$\hat{\delta}(x) = \left(\sum_i w_{i,x}\begin{bmatrix}1, & (X_i-x)' \\ X_i-x, & (X_i-x)(X_i-x)'\end{bmatrix}\right)^{-1}\sum_i w_{i,x}\begin{bmatrix}1 \\ X_i-x\end{bmatrix}Y_i \tag{14.12}$$

其中,$w_{i,x}=w((X_i-x)/R_x)$。

我们给出以下假定。

假定 14.3 (X_i,Y_i)独立同分布,$u_i=Y_i-g(X_i)$,$E(u_i^4)$是有限的。$g(x)$和 $f(x)$都是直到三阶可微,且 $\sigma^2(x)=E(u_i^2|X_i=x)$在 x 的范围里连续。

假定 14.4 当 $n\to\infty$,$k\to\infty$ 时,$k^{q+2}/n^2\to\infty$ 且 $k^{q+6}/n^6\to 0$。

接下来的定理给出了 $\hat{\delta}(x)$的渐近正态结果。

定理 14.5 在假定 14.3 和 14.4 下

$$D(n)\left(\hat{\delta}(x)-\delta(x)-\begin{bmatrix}(k/n)^{2/q}\mu_L(x) \\ 0\end{bmatrix}\right)\to N(0,\Sigma_x)\text{依分布收敛}$$

其中,$D(n)=\begin{bmatrix}k^{1/2}, & 0 \\ 0, & \mathrm{d}_n I_q\end{bmatrix}$,

$$\mathrm{d}_n = k^{1/2}(k/n)^{1/q}$$

$$\mu_L(x) = (1/2)\kappa_2\mathrm{tr}\{g^{(2)}(x)\}/(c_0 f(x))^{2/q}$$

$$\Sigma_x = \begin{bmatrix}c_0\kappa\sigma^2(x), & 0 \\ 0, & c_0\kappa_{22}\sigma^2(x)I_q/\kappa_2^2\end{bmatrix}$$

对定理 14.5 的证明在 14.9 节中给出。

$\hat{g}(x)$的渐近分布($\hat{\delta}(x)$的第一个元素)可以作为一个推论得到。

推论 14.1 在与定理 14.5 相同的条件下,我们有

$$k^{1/2}(\hat{g}(x)-g(x)-(k/n)^{2/q}\mu_L(x))\to N(0,c_0\kappa\sigma^2(x))\text{依分布收敛}$$

14.4 局部常数 k-nn 估计的交错鉴定

我们考虑非参数回归模型

$$Y_i = g(X_i) + u_i, \quad i=1,2,\cdots,n$$

以 X_i 为中心的 k-nn 距离定义如下：

定义 14.2　$R_i \equiv R_n(X_i) \stackrel{\text{def}}{=}$ 在所有的 $X_j(j \neq i)$ 中到 X_i 的 k-nn 欧几里得距离。

为了简化证明，我们对加权函数施加一个很强的条件。我们假定 $w(\cdot)$: $\mathbb{R}^q \to \mathbb{R}$ 是一个有界的非负加权函数，$w(v) = w(-v)$，$\int w(v)\,dv = 1$，且对 $\|v\| \geq 1$，$w(v) = 0$，其中 $\|v\|$ 表示 v 的欧几里得范数。可以证明这一章中讨论的所有结果保持不变，如果对 $\|v\| \geq 1$，$w(v) = 0$ 的条件放宽为 $\|v\| > 1$，$w(v) = 0$。但是，由于必须考虑一个附加项，证明过程将变长。注意当在证明中改变变量时，$\|v\| = 1$ 对应于第 k 个近邻，计算均方误差时阶数为 k^{-2}，而均方误差首项的阶数为 $k^{-1} + (k/n)^{2/q}$，小于 k^{-2}。因此，第 k 个近邻对均方误差的计算有一个渐近可忽略的贡献。因此，不管我们假定的是当 $\|v\| \geq 1$ 时，$w(v) = 0$ 还是 $\|v\| > 1$ 时，$w(v) = 0$，都不影响 k-nn 估计量的渐近分析。　423

令 $\hat{G}_k$ 表示 $n \times 1$ 维向量，它的第 i 个元素为 $\hat{g}(X_i)$。那么我们知道 $\hat{G}_k = M_n(k)Y$，其中 M_n 是一个 $n \times n$ 维矩阵，它的第(i,j)个元素由 $w_{ij} \Big/ \sum_{l=1}^{n} w_{il}$ 给出，其中 $w_{ij} = w((X_i - X_j)/R_i)$ 且 R_i 依赖于 k。

下面给出的是 Li(1987)研究的三个著名的选择 k 的步骤。

(i) Mallows 的 C_L(或 Cp)(Mallows(1973))：选择使以下目标函数最小化的 $\hat{k}$ 值：

$$\hat{k}_C = \arg\min_k n^{-1} \sum_{i=1}^{n} [Y_i - \hat{g}(X_i)]^2 + 2\sigma^2 \operatorname{tr}[M_n(k)]/n \tag{14.13}$$

其中 σ^2 是 u_i 的方差。在实践中，我们可以通过 $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \hat{u}_i^2$ 来估计 σ^2，其中 $\hat{u}_i = Y_i - \hat{g}(X_i)$。

(ii) 广义交错鉴定(Craven 和 Wahba(1979))：选择 $\hat{k}$ 以最小化

$$\hat{k}_{\mathrm{GCV}} = \arg\min_k \frac{n^{-1} \sum_{i=1}^{n} [Y_i - \hat{g}(X_i)]^2}{(1 - n^{-1}\operatorname{tr}[M_n(k)])^2} \tag{14.14}$$

(iii) 去一交错鉴定(Stone(1974))：选择 $\hat{k}$ 以最小化

$$\mathrm{CV}_k = \sum_{i=1}^{n} [Y_i - \hat{g}_{-i}(X_i)]^2 \tag{14.15}$$

其中 $\hat{g}_{-i}(X_i) = \sum_{j \neq i} Y_j w_{ij} / \sum_{j \neq i} w_{ij}$ ($w_{ij} = w((X_i - X_j)/R_i)$) 是 $g(X_i)$的去一 k-nn 估计值。

GCV 和 C_L方法的计算比去一交错鉴定方法的计算要简单。

Li(1987)证明了以上三种方法渐近等价。它们都在某种意义上得出了一个最优的平滑,即

$$\frac{\int[\hat{g}_{\hat{k}}(x)-g(x)]^2\mathrm{d}F(x)}{\inf_k\int[\hat{g}_k(x)-g(x)]^2\mathrm{d}F(x)}\xrightarrow{p}1 \tag{14.16}$$

其中式 $\hat{g}_k(x)=\hat{g}(x)$ 由式(14.7)定义,k 为一般取值,其中 $\hat{g}_{\hat{k}}(x)$ 是 $g(x)$ 的 k-nn 估计,其运用上述方法之一选择 k 值。也就是说,在式(14.6)中,$\hat{k}=\hat{k}_C$,或者 $\hat{k}=\hat{k}_{GCV}$,或者 $\hat{k}=\hat{k}_{CV}$。

424 方程(14.16)说明了以上三种方法的任何一种在某种意义上都能得到一个最佳的 k 值,使得最终的渐近加权积分平方误差等于最小的可能加权积分平方误差。

对数据驱动的 $\hat{k}$,要建立 $\hat{g}(x)$ 的渐近正态性必须得到一个非随机的 k,记为 k_0,使得 $\hat{k}/k_0\xrightarrow{p}1$。我们转向下面的去一交错鉴定方法的推导。

假定 14.5 (X_i,Y_i)独立同分布,$u_i=Y_i-g(X_i)$,$E(u_i^4)$是有限的,$g(x)$和 $f(x)$都连续且直到三阶可微。$\sigma^2(x)=E(u_i^2|X_i=x)$是关于 x 连续的。令$\mathcal{S}$表示 $M(\cdot)$的支撑集,则对于某些 $\delta>0$,有 $\inf_{x\in\mathcal{S}}f(x)\geqslant\delta$。

假定 14.6 $w(\cdot)$是一个有界、对称、非负函数,当 $\|v\|\geqslant1$ 时,$w(v)=0$。$\int w(v)\mathrm{d}v=\int_{\|v\|\leqslant1}w(v)\mathrm{d}v=1$,$\int w(v)vv'\mathrm{d}v=\kappa_2I_q$,$\int w^2(v)\mathrm{d}v=\kappa$,$\int w^2(v)vv'\mathrm{d}v=\kappa_{22}I_q$,其中 I_q 是一个 q 维单位矩阵。κ_2、κ、和 κ_{22}都是有限的正常数。

假定 14.7 对于任意小的 $\varepsilon\in\left(0,\frac{1}{2}\right)$,有 $k\in\Lambda=[n^{\varepsilon},n^{1-\varepsilon}]$。

条件 $\|v\|\geqslant1$ 时,$w(v)=0$ 可以放宽为 $\|v\|>1$ 时,$w(v)=0$,结论不会改变。但是这需要更长的论证。假定 14.7 意味着当 $k/n\to0$ 且当 $n\to\infty$ 时,$k\to\infty$。它保证了当样本大小增加时,k-nn 估计值的偏差和方差收敛于零。

以 $g^{(1)}(x)$表示 $g(x)$的 $q\times1$ 维一阶导数向量,$g^{(2)}(x)$表示 $g(x)$的 $q\times q$ 维二阶导数矩阵,令 $g_s(x)$($f_s(x)$)表示 $g(\cdot)$($f(\cdot)$)对 x_s 的偏导,其中 x_s 为 x 的第 s 个分量($s=1,\cdots,q$)。

Ouyang, Li 和 Li(2006)证明了式(14.15)定义的 $\mathrm{CV}_k(k)$的首项具有形式 $\Phi_1(k/n)^{4/q}+\Phi_2k^{-1}$,在 $k\in\Lambda$ 是均匀的,其中,

$$\Phi_1=c_0^{-4/q}\kappa_2^2\int\left(\left[\frac{1}{2}f(x)\mathrm{tr}\left[{}^{(2)}g(x)\right]+\sum_{s=1}^{q}f_s(x)g_s(x)\right]\right)^2\times[f(x)]^{-(q+4)/q}M(x)\mathrm{d}x \tag{14.17}$$

$$\Phi_2 = c_0\kappa\int\sigma^2(x)M(x)f(x)\mathrm{d}x \tag{14.18}$$

因此,我们有

425

$$\mathrm{CV}_{kC}(k) = \Phi_1(k/n)^{4/q} + \Phi_2 k^{-1} + o((k/n)^{4/q} + k^{-1}) \tag{14.19}$$

在 $k\in\Lambda$ 是均匀的。令 k_0 表示最小化 $\mathrm{CV}_k(k)$ 的首项,$\Phi_1(k/n)^{4/q}+\Phi_2 k^{-1}$ 的 k 值。很容易证明 $k_0=a_0 n^{4/(4+q)}$,其中 $a_0=[q\Phi_2/(4\Phi_1)]^{q/(4+q)}$。之前提到 $\hat{k}$ 为 k 的交错鉴定值,由式(14.19)立即有下述结论。

定理 14.6 在假定 14.5 至 14.7 的情况下,我们有

$$\hat{k} = k_0 + o_p(k_0) \quad 或等价地, \quad \hat{k}/k_0 \xrightarrow{p} 1$$

Ouyang 等对定理 14.6 进行了证明。

令 $\hat{g}_{\hat{k}}(x)$ 表示 $k=\hat{k}$ 时的 $\hat{g}(x)$。下一个定理给出了 $\hat{g}_{\hat{k}}(x)$ 的渐近分布。

定理 14.7 在与定理 14.6 相同的条件下,对于 $x\in\mathbb{R}^q, f(x)>0$,我们有

$$\hat{k}^{1/2}(\hat{g}_{\hat{k}}(x) - g(x) - (\hat{k}/n)^{2/q}\mu L(x)) \to N(0, c_0\kappa\sigma^2(x))\text{依分布收敛}$$

其中

$$\mu L(x) = (1/2)\kappa_2\left\{\sum_{s=1}^{q} f_s(x)g_s(x) + f(x)\mathrm{tr}\left[\overset{(2)}{g}(x)\right]\right\}\Big/\left[c_0^{q/2}f(x)^{(q+2)/2}\right]$$

定理 14.7 的证明遵循定理 14.4、定理 14.6 及随机同等连续性定理。详细证明过程参见 Ichimura(2000)。

14.5 局部线性 k-nn 估计的交错鉴定

非参数回归模型与之前考虑的相同,即

$$Y_i = g(X_i) + u_i, \quad i = 1,\cdots,n$$

之前提到 $\delta(X_i)=(g(X_i), g^{(1)}(X_i)')'$。$\delta(X_i)$ 的去一 k-nn 估计值由式(14.20)给出:

426

$$\hat{\delta}_{-i}(X_i) = \left[\sum_{j\neq i}^{n} wR_{i,ij}\begin{pmatrix}1, & (X_j - X_i)' \\ X_j - X_i, & (X_j - X_i)(X_j - X_i)'\end{pmatrix}\right]^{-1} \times \sum_{j\neq i}^{n} wR_{i,ij}\begin{bmatrix}1 \\ X_j - X_i\end{bmatrix}Y_j, \tag{14.20}$$

其中 $wR_{i,ij}=R_i^{-q}w((X_j-X_i)/R_i)$。

$g(X_i)$ 的去一 k-nn 估计值由 $\hat{g}_{-i,L}(X_i)=e_1'\hat{\delta}_{-i}(X_i)$ 给出,其中 e_1 为 $q\times 1$ 向量,它的第一个元素为 1,e_1 向量的其他元素都是 0。选择 k 值来最小化交错鉴定目标函数

$$\mathrm{CV}_L(k) = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{g}_{-i,L}(X_i))^2 M(X_i) \tag{14.21}$$

其中 $\hat{g}_{-i,L}(X_i) = e_1'\hat{\delta}_{-i}(X_i)$，$M(\cdot)$为一个加权函数。

Ouyang 等证明了，$\mathrm{CV}_{kL}(k) = \Phi_{1,L}\left(\frac{k}{n}\right)^{4/q} + \Phi_2 k^{-1} + o_p\left(\left(\frac{k}{n}\right)^{4/q} + k^{-1}\right)$在 $k \in \Lambda$ 是均匀的。

其中 Φ_2 的定义与式(14.18)中的相同，且

$$\Phi_{1,L} = c_0^{-4/q}\kappa_2^2\int\left(\frac{1}{2}f(x)\mathrm{tr}[g^{(2)}(x)]\right)^2[f(x)]^{-(q+4)/q}M(x)\mathrm{d}x \tag{14.22}$$

我们观察到，与逐点结果相似，局部线性交错鉴定目标函数的偏误首项有一个简单的形式。

令 $k_{0,L}$表示最小化 $\mathrm{CV}_k(k)$的首项 $\Phi_{1,L}(k/n)^{4/q} + \Phi_2 k^{-1}$的 k 值，很容易证明 $k_{0,L} = a_L n^{4/(4+q)}$，其中 $a_L = [q\Phi_2/(4\Phi_{1,L})]^{q/(4+q)}$ 是一个正常数。下一个定理证明了交错鉴定值 k 渐近等于 $k_{0,L}$。

定理 14.8 在假定 14.5 至假定 14.7 条件下，令 $\hat{k}_L$ 表示 k 的局部线性交错鉴定值，则

$$\hat{k}_L = k_{0,L} + o(k_{0,L}) \quad \text{或等价地，} \quad \hat{k}_L/k_{0,L} \xrightarrow{p} 1$$

427

除了 $k = \hat{k}_L$，$\hat{\delta}_{\hat{k}}(x)$的定义类似于 $\hat{\delta}(x)$；$\hat{\delta}_{\hat{k}}(x)$的渐近分布由下述定理给出。

定理 14.9 在与定理 14.8 相同的条件下，

$$\hat{D}(n)\left(\hat{\delta}_{\hat{k}}(x) - \delta(x) - \begin{bmatrix}(\hat{k}/n)^{2/d}\mu_L(x)\\ 0\end{bmatrix}\right) \to N\left(0, \sum_x\right) \text{ 依分布收敛}$$

其中

$$\hat{D}(n) = \begin{bmatrix}\hat{k}_L^{1/2}, 0\\ 0, \hat{d}_n I_q\end{bmatrix}, \quad \hat{d}_n = \hat{k}_L^{1/2}(\hat{k}_L/n)^{1/q}$$

$$\mu_L(x) = (1/2)\kappa_2\mathrm{tr}\{g^{(2)}(x)\}/(c_0 f(x))^{q/2}$$

$$\sum_x = \begin{bmatrix}c_0\kappa\sigma^2(x), & 0\\ 0, & c_0\kappa_{22}\sigma^2(x)I_q/\kappa_2^2\end{bmatrix}$$

由定理 14.8 和与定理 14.7 证明过程中类似的论证，可证明定理 14.9。故在此省略。

$\hat{\delta}_{\hat{k}}(x)$第一个元素 $\hat{g}_{L,\hat{k}}(x)$的渐近分布作为定理 14.9 的一个推论给出。

推论 14.2 在与定理 14.9 相同的条件下，有

$$\hat{k}_L^{1/2}(\hat{g}_{L,\hat{k}}(x) - g(x) - (\hat{k}_L/n)^{2/q}\mu_L(x)) \xrightarrow{d} N(0, c_0\kappa\sigma^2(x))$$

14.6　k-nn 方法估计半参数模型

我们之前讨论的使用核方法的半参数模型都可以采用 k-nn 方法来估计。例如，关于使用 k-nn 估计方法，Newey(1994a)考虑了平均导数的估计，Liu 和 Lu (1994a)研究了部分线性模型的估计，而 Yatchew，Sun 和 Deri(2003)估计了单(多)指数模型。在这一节，我们讨论一个流行的半参数模型的估计，即对第 7 章考虑的部分线性模型进行估计。

部分线性模型形式如下：　　428

$$Y_i = X_i'\beta + \theta(Z_i) + u_i \tag{14.23}$$

其中 $Z_i \in \mathbb{R}^q$ 是连续随机变量的一个向量。

类似于 β 的基于核估计值推导的情形，可以通过使用下式的基于 k-nn 的方法来估计 β：

$$\hat{\beta} = S_{X-\hat{X}}^{-1} S_{X-\hat{X},Y-\hat{Y}} \tag{14.24}$$

其中 $S_{X-\hat{X}} = n^{-1}\sum_i (X_i - \hat{X}_i)(X_i - \hat{X}_i)'$，$S_{X-\hat{X},Y-\hat{Y}} = n^{-1}\sum_i (X_i - \hat{X}_i)(Y_i - \hat{Y}_i)I_i$，$I_i = \mathbf{1}(\hat{f}_i > b)$，$b = b_n(>0)$ 是一个修剪参数，$\hat{X}_i = \sum_{j=1}^n X_j w_{ij} / \sum_{j=1}^n w_{ij}$，$\hat{Y}_i = \sum_{j=1}^n Y_j w_{ij} / \sum_{j=1}^n w_{ij}$，$\hat{f}_i = (nR_i^q)^{-1}\sum_{j=1}^n w_{ij}$，$w_{ij} = w((X_i - X_j)/R_i)$，且 $\hat{X}_i$、$\hat{Y}_i$ 和 $\hat{f}_i$ 分别是 $E(X_i|Z_i)$、$E(Y_i|Z_i)$ 和 $f(X_i)$ 的 k-nn 估计值。

条件 14.1　w 是一个 ν 阶加权函数(即一个 ν 阶核函数)；当 $n\to\infty$ 时，$k\to\infty$，$b\to 0$，$n^{-1}k^2b^4\to\infty$，$n^{q-4}k^{4-2q}b^{-4q}\to 0$，$k^{4\nu}n^{q-4\nu}b^{-4q}\to 0$ 且 $k^4n^{-4}b^{-4q}\to 0$。

定理 14.10　与定理 7.1(核方法)以及上面的条件 14.1 给出的条件相同，我们有

$$\sqrt{n}(\hat{\beta} - \beta) \to N(0, V) \text{ 依分布收敛}$$

其中，$V = E[\tilde{X}_i\tilde{X}_i']\{E[\sigma^2(X_i,Z_i)\tilde{X}_i\tilde{X}_i']\}^{-1}E[\tilde{X}_i\tilde{X}_i']$，$\tilde{X}_i = X_i - E(X_i|Z_i)$，$\sigma^2(X_i,Z_i) = E(\mu_i^2|X_i,Z_i)$。

有关定理 14.10 的大致证明过程在 14.9 节给出。

定理 14.10 表明基于 k-nn 的半参数估计值 $\hat{\beta}$ 有着与基于核的估计值相同的渐近分布。可以参见 Newey(1994a)的结论。

14.7 k-nn 方法的模型设定检验

关于构造一致模型设定检验有大量的文献,诸如在第 12 章中提到的,不过大部分现有的检验是基于核、序列、样条或者小波方法。而 k-nn 方法是最流行的非参数估计技术之一,我们知道的仅有的基于 k-nn 的检验是 Stute 和 Manteiga(1996)以及 Li(2006)。Stute 和 Manteiga 考虑了一个单变量回归模型,他们运用一个单调变换来使他们的 k-nn 检验等价于一个核检验。而 Li 提出了一个对一般多元回归模型的检验。

429

考虑一个参数回归模型,形式如下:

$$Y_i = m(X_i, \beta) + \mu_i \tag{14.25}$$

其中 $X \in \mathbb{R}^q$,β 是未知参数的一个 $p \times 1$ 维向量,我们感兴趣的是检验对式(14.25)的正确设定的原假定,即对于某些 $\beta \in \mathcal{B} \subset \mathbb{R}^p$,$H_0^a$:几乎处处 $E(Y_i | X_i = x) = m(x, \beta)$,这里$\mathcal{B}$是$\mathbb{R}^p$的一个紧凑子集。给定 $u_i = Y_i - m(X_i, \beta)$,原假定可以等价地写成

$$H_0: E(u_i \mid X_i = x) = 0 \text{ a.e.},\text{某些} \beta \in \mathcal{B}$$

备择假定与 H_0 相反,即 $H_1: E(u_i | X_i = x) \neq 0$,对在正测度集上的任何 $\beta \in \mathcal{B}$。我们在 $I = E[u_i E(u_i | X_i) f(X_i)]$ 的基础上构造检验统计量。I 的相应样本形式为 $I_n = \frac{1}{n}\sum_i u_i E(u_i \mid X_i) f(X_i)$。为了得到一个可行的检验统计量,我们通过 $\hat{u}_i = Y_i - m(X_i, \hat{\beta})$ 来估计 u_i,其中 $\hat{\beta}$ 是在 H_0 假定下 β 的任意$\sqrt{n}$一致估计量,即基于式(14.25)的非线性最小二乘估计量 β。

首先,我们定义 R_i 的简化符号如下:

$R_i = R_n(X_i) = X_i$ 与第 k 个近邻的欧几里得距离

接着我们通过以下去一 k-nn 估计量来估计 $E(\varepsilon_i | X_i)$ 和 $f(X_i)$:

$$\hat{E}_{-i}(u_i \mid X_i) = \frac{1}{n-1}\sum_{j \neq i}^{n} \frac{1}{R_i^q}\hat{u}_j w\left(\frac{X_i - X_j}{R_i}\right) / \hat{f}_{-i}(X_i)$$

$$\hat{f}_{-i}(X_i) = \frac{1}{n-1}\sum_{j \neq i}^{n} \frac{1}{R_i^q} w\left(\frac{X_i - X_j}{R_i}\right) \tag{14.26}$$

在 I_n 的表达式中以 $\hat{u}_i$、$\hat{E}_{-i}(u_i | X_i)$ 和 $\hat{f}_{-i}(X_i)$ 分别取代 u_i、$E_i(u_i | X_i)$ 和 $f_i(X_i)$ 得到一个可行的检验统计量:

$$\hat{I}_n = \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j \neq i}^{n} \frac{1}{R_i^q}\hat{u}_i \hat{u}_j w_{ij} \tag{14.27}$$

其中 $w_{ij}=w((X_i-X_j)/R_i)$。注意因为 $R_i\neq R_j$，有 $w_{ij}\neq w_{ji}$。 430

当 $\|v\|>1$ 时，$w(v)=0$ 这个条件可以用条件 $\int w(v)\|v\|^4 dv<\infty$ 取代，不会改变这一节的结论。例如，在实践中可以使用一个标准正态核，尽管证明过程将会更加繁琐。

在与定理 12.1(采用核方法)相同的条件下，并在假定 14.1 和 14.2 下，Li(2006)证明了 $\hat{I}_n$ 的渐近原分布由它的首项 $\hat{I}_{n1}$ 决定，其中 $\hat{I}_{n1}$ 由下式给出：

$$\hat{I}_{n1}=\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{1}{2}u_iu_j\left(\frac{w_{ij}}{R_i^q}+\frac{w_{ij}}{R_j^q}\right)$$

$$\equiv\frac{2}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}^{n}H_{1n}(Z_i,Z_j) \tag{14.28}$$

其中 $H_{1n}(Z_i,Z_j)=(1/2)u_iu_j((w_{ij}/R_i^q)+(w_{ji}/R_j^q))$，而 $Z_i=(u_i,X_i,R_i)$。$E(u_i|X_i)=0$意味着 $E[H_{1n}(Z_i,Z_j)|Z_i]=0$。因此，$\hat{I}_{n1}$是一个二阶退化 U 统计量。$\hat{I}_{n1}$的渐近分布可从关于退化 U 统计量的 Hall(1984)中心极限定理推导出来。

Li(2006)证明了在假定 H_0^a 下，有 $\hat{I}_{n1}=O_p((nk)^{-1/2})$ 和 $\hat{I}_n-\hat{I}_{n1}=O_p(n^{-1})=o_p((nk)^{-1/2})$。因此，在假定 H_0^a 下，$\hat{I}_{n1}$是 $\hat{I}_n$ 的首项，即 $\hat{I}_{n1}$和 $\hat{I}_n$ 有相同的渐近分布。定义

$$\hat{\sigma}_n^2=\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j>i}^{n}\hat{u}_i^2\hat{u}_j^2(w_{ij}/R_i^q+w_{ji}/R_j^q)^2 \tag{14.29}$$

Li(2006)证明了下述结论。

定理 14.11 在引理 A.15 和 A.16 以及 H_0 假定下，有

$$\hat{T}_n\stackrel{\text{def}}{=}n\hat{I}_n/\{\hat{\sigma}_n^2\}^{1/2}$$

$$\equiv(nk)^{1/2}\hat{I}_n/\{(k/n)\hat{\sigma}_n^2\}^{1/2}\xrightarrow{d}N(0,1)$$

可以证明，无论是在假定 H_0 还是 H_1 下，都有$(k/n)\hat{\sigma}_n^2=O_p(1)$。当原假定为真时，$\hat{I}_n$ 的首项是一个以$(nk)^{-1/2}$的速率收敛于零的退化 U 统计量。可以证明，当 H_1 为真时，$\hat{I}_n$ 收敛于一个正常数，记为 $C>0$。因此，在 H_1 下，$\hat{T}_n$ 以$(nk)^{1/2}$的速率发散至无穷。所以，基于 $\hat{T}_n$ 的检验是一个一致性检验。

14.7.1 自举法检验

431

可以证明，在假定 H_0 下，$\hat{T}_n$ 检验以一个相当低的速率 $O_p((k/n)^{1/2})$收敛于它的渐近标准正态分布。这个速率与选择 $k/n\sim h^q$ 的基于核方法检验的速率

相同。其中,h 是核检验的平滑参数。各种自举法被使用以更好地逼近大量现有的非参数检验的零分布;参见 Härdle 和 Mammen(1993),Li 和 Wang(1998),以及其中的参考文献。

也可以通过原始自举法逼近 $\hat{T}_n$ 的零分布。令 u_i^* 表示满足下列条件的自举误差(基于 $\hat{u}_i$):

$$(\mathrm{i})\ E^*(u_i^*)=0,\quad (\mathrm{ii})\ E^*(u_i^{*2})=\hat{u}_i^2,\quad (\mathrm{iii})\ E^*(u_i^{*3})=\hat{u}_i^3 \tag{14.30}$$

其中 $E^*(\cdot)=E(\cdot|\chi_n)$,$\chi_n=\{(X_i,Y_i)\}_{i=1}^n$。例如,由 $u_i^*=a\hat{u}_i$ 和 $u_i^*=b\hat{u}_i$ 得到两点原始自举法残差值。$u_i^*=a\hat{u}_i$ 的概率为 $r=(\sqrt{5}+1)/(2\sqrt{5})$,而 $u_i^*=b\hat{u}_i$ 的概率为 $1-r$,其中 $a=(1-\sqrt{5})/2$,$b=(1+\sqrt{5})/2$,满足式(14.30)给出的三个条件。

接下来,生成 $Y_i^*=m(X_i,\hat{\beta})+u_i^*$,并且使用基于原模型 $Y_i^*=m(X_i,\beta)+\text{error}$ 的自举样本 $\{X_i,Y_i^*\}_{i=1}^n$,通过非线性最小二乘法来估计 β。令 $\hat{\beta}^*$ 表示得到的估计值。自举残差由 $\hat{u}_i^*=Y_i^*-m(X_i,\hat{\beta}^*)$ 得到。$\hat{T}_n^*$ 的定义类似于 $\hat{T}_n$,只是将式子中的 $\hat{u}_i$ 用 $\hat{u}_i^*$ 取代。在实践中重复以上过程多次,比如 $B=399$ 次,得到 B 的自举法统计量 $\{\hat{T}_{n,j}^*\}_{j=1}^B$。如果 $\hat{T}_n>\hat{T}_{n,(\alpha)}^*$,则在显著性水平 α 上拒绝原假定 H_0,其中 $\hat{T}_{n,(\alpha)}^*$ 是 $\{\hat{T}_{n,j}^*\}_{j=1}^B$ 的上 α 分位点。

在 Li(2006)中报告的模拟表明,基于 k-nn 的检验与基于核估计的检验很相似;参见 Härdle 和 Mammen(1993),Li 和 Wang(1998)。特别地,如果采用渐近正态临界值,则 $\hat{T}_n$ 检验通常显著性不足,然而,自举法检验运行良好。因此,我们建议在实际中采用自举法。

以上给出的基于 k-nn 检验的结论可以很好地扩展到对弱相依时间序列模型中的正确参数回归函数形式(运用 k-nn 方法)的检验情形。在此情形中,我们可以对弱相依数据使用一个关于衰退 U 统计量的中心极限定理(如 Fan 和 Li(1999a),来证明对弱相依数据这个检验有相同的渐近分布。我们也可以扩展这些结论来检验使用 k-nn 方法的更加复杂的原模型。例如,一个非参数遗漏变量检验,单个指数模型检验,或者部分线性模型对普通非参数模型的检验;参见 Fan 和 Li(1999a),Delgado 和 Manteiga(2001)。

432

相关的工作还有 Delgado 和 Stengos(1994),他们运用 k-nn 方法构造了一个模型设定检验,对参数原模型针对非嵌套的非参数备择模型进行检验。Robinson(1987)在条件异方差误差形式未知的线性回归模型中构造了一个基于 k-nn 估计的自适应(有效)估计方法。

14.8　对于 x 的不同的分量采用不同的 k 值

到现在为止,我们只讨论了使用一个标量 k 的情况。在实践中,我们可能用到乘积核 $W((X_i - x)/R_x) = \prod_{s=1}^{q} w((X_{is} - X_s)/R_{xs})$,其中 R_{xs}表示在 X_{is}中 x_s 的第 k_s 个近邻与 x_s 的距离,$i=1,\cdots,n,s=1,\cdots,q$。这个方法对于不同的 x_s 允许使用不同的 k_s(类似于核估计中对于不同的 x_s 使用不同的 h_s)并且期望得到更准确的有限样本估计。在实践中,我们也可以使用 14.4 节中讨论的 Mallow 的 C_L 方法、CV 方法和 GCV 方法来选择 k_s。正如预期的,与使用标量 k 相比,对于 x_s 的不同分量使用不同的 k_s 产生了更准确的非参数估计(与使用的欧几里得距离有关)。但是,当允许 k_s 在协变量间不同时,它的渐近分析相当复杂而且目前这种情况在理论上是无效的。

14.9　证　明

因为 n 和 $n-1$ 都产生了一致的渐近结果,所以我们对 n 和 $n-1$ 不加以区分。例如,我们可以用 $n^{-2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}E(A_{ij}) = E(A_{12})$ 取代$[n(n-1)/n^2]E(A_{12})$来简化我们的分析。

令 $S_r=\{v: \| v-x \| < r\}$(以 x 为圆心,r 为半径的球),$G(r)=P[X_i \in S_r]$,$S_x=\{v: \| v-x \| < R_x\}$,$P(S_x)=P[X_i\in S_x]$,显然有 $G(R_x)=P(S_x)$。

引理 14.1　令 $\Phi(r)=\frac{1}{r^{\lambda}G^{\gamma}(r)}$,令 λ、γ 为整数使得 $E[\Phi(R_i)]$存在,则　433

$$E[\Phi(R_i) \mid X_i] = (c_0 f(X_i))^{\lambda/q}\left(\frac{k}{n}\right)^{-\lambda/q-\gamma} + (\text{s. o.})$$

其中 $c_0=\pi^{q/2}/\Gamma((q+2)/2)$是$\mathbb{R}^q$集里单位球的体积,并且(s. o.)表示概率阶数(probability order)小于$(k/n)^{-\lambda/q-\gamma}$的项。

证明:运用 Mack 和 Rosenblatt(1979)中的等式(12)以及 Liu 和 Lu(1997)中的引理 1 可以证明,对于 $m=(\lambda+\eta)/q$,其中 m 为整数,η 为小于或等于 $q-1$ 的非负整数($0\leqslant\eta\leqslant q-1$),有

$$E[\Phi(R_i) \mid X_i] = (c_0 f(X_i))^{\lambda/q}\frac{n!(k-m-\gamma)!}{k!(n-m-\gamma)!}\left(\frac{k-m-\gamma}{n-m-\gamma}\right)^{\eta/q} + (\text{s. o.}) \tag{14.31}$$

注意

$$\frac{n!}{k!}\frac{(k-m-\gamma)!}{(n-m-\gamma)!}\left(\frac{k-m-\gamma}{n-m-\gamma}\right)^{\eta/q}=\left(\frac{n}{k}\right)^{m+\gamma}\left(\frac{k}{n}\right)^{\eta/q}+(\mathrm{s.o.})$$
$$=\left(\frac{k}{n}\right)^{\eta/q-m-\gamma}+(\mathrm{s.o.})$$
$$=\left(\frac{k}{n}\right)^{-\lambda/q-\gamma}+(\mathrm{s.o.})$$

将此式代入式(14.31)证明了引理14.1。 □

注意引理14.1意味着 $R_i \sim (c_0 f(X_i))^{-1/q}(k/n)^{1/q}(r=0,\lambda=-1)$。这是一个期望中的结果,因为 R_i 相当于核估计中的平滑参数 h 且 $h\sim(k/n)^{1/q}$。引理14.1也意味着 $G(R_i)\sim(k/n)$。

引理14.2 令 $A(x)$ 为 x 的可测函数,则

$$E\left[A(X_j)w\left(\frac{X_j-X_i}{R_i}\right)\mid X_i,R_i\right]=\frac{k-1}{n}\frac{1}{G(R_i)}\times\int_{\|X_j-X_i\|<R_i}f(X_j)A(X_j)w\left(\frac{X_j-X_i}{R_i}\right)\mathrm{d}X_j \tag{14.32}$$

证明:这个证明过程直接由 Mack 和 Rosenblatt(1979,式(22))以及对 $\|X_j-X_i\|\geqslant R_i$ 有 $w((X_j-X_i)/R_i)=0$ 的事实可得。 □

434

我们知道

$$\begin{aligned}G(R_i)&=P[\|X_j-X_i\|<R_i\mid R_i]\\&=\int_{\|x_j-x_i\|<R_j}f(x_i\mid R_i)f(x_j)\mathrm{d}x_i\mathrm{d}x_j\\&=\int f(x_i\mid R_i)P[\|X_j-x_i\|<R_i\mid x_i,R_i]\mathrm{d}x_i\\&=\int f(x_i\mid R_i)(k/n)\mathrm{d}x_i\\&=k/n\end{aligned}$$

由引理14.1我们知道可以用 k/n 取代 $G(R_i)$。如果在式(14.32)中以 k/n 取代 $G(R_i)$,将得到

$$E\left[A(X_j)w\left(\frac{X_j-X_i}{R_i}\right)\mid X_i,R_i\right]=\int_{\|x_i-X_i\|<R_i}f(x_j)A(x_j)\times w\left(\frac{x_j-X_i}{R_i}\right)\mathrm{d}x_j \tag{14.33}$$

另一方面,我们知道

$$E\left[A(X_j)w\left(\frac{X_j-Xx_i}{R_i}\right)\mid X_i,R_i\right]=\int_{\|x_j-X_i\|<R_i}f(x_j\mid R_i)A(x_j)\times w\left(\frac{x_j-X_i}{R_i}\right)\mathrm{d}x_j \tag{14.34}$$

对比式(14.33)与式(14.34)我们得到

$$\int_{\| x_j-X_i \| < R_i} f(x_j \mid R_i) A(x_j) w\left(\frac{x_j - X_i}{R_i}\right) \mathrm{d}x_j$$

$$= \int_{\| x_j-x_i \| < R_i} f(x_j) A(x_j) w\left(\frac{x_j - X_i}{R_i}\right) \mathrm{d}x_j \tag{14.35}$$

方程(14.35)表明了在计算中我们可以用$f(x_j)$有效地取代$f(x_j \mid R_i)$。

14.9.1 定理 14.1 的证明

435

证明:运用式(14.33)我们有

$$E[\hat{f}(x) \mid R_x] = E\left[R_x^{-q} w\left(\frac{X_1 - x}{R_x}\right) \mid R_x\right]$$

$$= R_x^{-q} \int w\left(\frac{x_1 - x}{R_x}\right) f(x_1)\, \mathrm{d}x_1$$

$$= \int w(v) f(x + vR_x)\, \mathrm{d}v$$

$$= f(x) + (1/2) R_x^2 \int w(v) v' f^{(2)}(x) v \mathrm{d}v + O(R_x^3)$$

因此,$E(\hat{f}(x)) = f(x) + (1/2)\kappa_2 \mathrm{tr}[f^{(2)}(x)] E(R_x^2) + O(E(R_x^3))$。我们现在计算 $E(R_x^2)$。运用引理 14.1,有 $\lambda = -2, \gamma = 0$,我们得到

$$E(R_x^2) = (c_0 f(x))^{-2/q} (k/n)^{2/q} + (\text{s.o.}) \tag{14.36}$$

类似地,我们可以证明 $E(R_x^3) = O((k/n)^{3/q})$。因此,我们有 bias $\hat{f}(x) = (1/2)\kappa_2 \mathrm{tr}[f^{(2)}(x)](cf(x))^{-2/q}(k/n)^{2/q} + O((k/n)^{3/q})$。

接下来,

$$\mathrm{var}(\hat{f}(x) \mid R_x) = (nR_x^{2q})^{-1} \left\{ E\left[w\left(\frac{X_1 - x}{R_x}\right)^2 \mid R_x \right] + (\text{s.o.}) \right\}$$

$$= (nR_x^q)^{-1} f(x) \int w(v)^2 \mathrm{d}v + (\text{s.o.})$$

同时,由引理 14.1 有 $E(R_x^{-q}) = c_0 f(x)(n/k) + o(n/k)(\lambda = q, \gamma = 0)$。因此,$\mathrm{var}(\hat{f}(x)) = k^{-1} c_2 f(x)^2 + o(k^{-1})(c_2 = c\int w(v)^2 \mathrm{d}v)$。 □

14.9.2 定理 14.5 的证明

证明:类似于定理 2.7 的证明(核方法情形),在式(14.12)中引入一个单位矩阵 $I_{q+1} = \mathcal{G}_n^{-1}\mathcal{G}_n$,其中$\mathcal{G}_n = (nR_x^{q+2})^{-1}\begin{bmatrix} R_x^2, & 0 \\ 0, & I_q \end{bmatrix}$,我们得到

$$\hat{\delta}(x) = \delta(x) + [A^{1,x}]^{-1}\{A^{2,x} + A^{3,x}\} + (\text{s.o.})$$

436 其中

$$A^{1,x} = \frac{1}{nR^{q+2}} \sum_i w_{i,x} \begin{bmatrix} R_x^2, & R_x^2 (X_i - x)' \\ X_i - x, & (X_i - x)(X_i - x)' \end{bmatrix}$$

$$= \begin{bmatrix} A_{11}^{1,x}, & A_{12}^{1,x} \\ A_{21}^{1,x}, & A_{22}^{1,x} \end{bmatrix} \tag{14.37}$$

$$A^{2,x} = \frac{1}{nR_x^{q+2}} \sum_i w_{i,x} \begin{bmatrix} R_x^2 \\ X_i - x \end{bmatrix} (x_i - x)' g^{(2)}(x)(X_i - x)/2$$

$$= \begin{bmatrix} A_1^{2,x} \\ A_2^{2,x} \end{bmatrix} \tag{14.38}$$

$$A^{3,x} = \frac{1}{nR_x^{q+2}} \sum_i w_{i,x} \begin{bmatrix} R_x^2 \\ Xx_i - x \end{bmatrix} u_i = \begin{bmatrix} A_1^{3,x} \\ A_2^{3,x} \end{bmatrix} \tag{14.39}$$

引理 14.3 在定理 14.5 的条件下,我们有

(i) $A_{11}^{1,x} = f(x) + o_p(1)$,

(ii) $A_{21}^{1,x} = \kappa_2 f^{(1)}(x) + o_p(1)$,

(iii) $A_{12}^{1,x} = O_p((k/n)^{2/q})$,

(iv) $A_{22}^{1,x} = \kappa_2 f(x) I_q + o_p(1)$,

(v) $A^{1,x} = M + o_p(1)$,其中 $M = \begin{bmatrix} f(x), & 0 \\ \kappa_2 f^{(1)}(x), & \kappa_2 f(x) I_q \end{bmatrix}$。

(i)的证明:$A_{11}^{1,x} = \hat{f}(x)$是 k-nn 密度估计值,$\hat{f}(x) = f(x) + o_p(1)$在定理 14.1 中已证。 □

437 (ii)的证明:通过式(14.33)我们有

$$E[A_{21}^{1,x}] = E\left[R_x^{-(q+2)} w\left(\frac{X_1 - x}{R_x}\right)(X_1 - x)\right]$$

$$= E\left\{E\left[R_x^{-(q+2)} w\left(\frac{X_1 - x}{R_x}\right)(X_1 - x) \mid R_x\right]\right\}$$

$$= E\left\{R_x^{-1} \int w(v) v f(x + vR_x) \mathrm{d}v\right\}$$

$$= E\left\{R_x^{-1} \int w(v) v [f(x) + v' f^{(1)}(x) R_x] \mathrm{d}v\right\} + (s.o.)$$

$$= \kappa_2 f^{(1)}(x) + O((k/n)^{2/q}) \text{ 因为 } (s.o.) = O(E(R_x^2))$$

$$= O((k/n)^{2/q})$$

习题 14.1 证明了 $\mathrm{var}(A_{21}^{1,x})=O(n^{-1})$。因此，

$$A_{21}^{1,x}=E(A_{21}^{1,x})+O_p(n^{-1/2})=\kappa_2 f^{(1)}(x)+O_p((k/n)^{2/q}+n^{-1/2})\qquad\square$$

(iii)的证明：由引理 14.1 和公式(14.36)我们有 $E[R_x^2]=O((k/n)^{2/q})$。因此，由引理 14.3(ii)中 $A_{12}^{1,x}=O_p(1)$ 推出 $A_{12}^{1,x}=R_x^2(A_{21}^{1,x})'=O_p((k/n)^{2/q})O_p(1)$。

$\square$

(iv)的证明：使用式(14.33)我们有

$$\begin{aligned}E[A_{22}^{1,x}\mid R_x]&=E\left[R^{-(q+2)}(X_i-x)(X_i-x)'w\left(\frac{X_i-x}{R_x}\right)\mid R_x\right]\\&=R_x^{-(q+2)}\int(x_1-x)(x_1-x)'w\left(\frac{x_1-x}{R_x}\right)f(x_1)\mathrm{d}x_1\\&=\int vv'w(v)f(x)\mathrm{d}v+O(R_x^2)\end{aligned}$$

因此，由引理 14.1 得到

$$E[A_{22}^{1,x}]=E\{E[A_{22}^{1,x}\mid R_x]\}=\kappa_2 f(x)I_q+O((k/n)^{2/q})$$

根据在(ii)的证明中相同的论证，可以证明 $E[\|A_{22}^{1,x}-\kappa_2 f(x)I_q\|^2]=o(1)$。因此，$A_{22}^{1,x}=\kappa_2 f(x)I_q+o_p(1)$。　$\square$

(v)的证明：$A^{1,x}=M+o_p(1)$ 由上述(i)—(iv)得到。　$\square$

引理 14.4

(i) $A_1^{2,x}=(k/n)^{2/q}[f(x)\mu_L(x)+o_p(1)]$

(ii) $A_2^{2,x}=O_p((k/n)^{2/q})$　438

(i)的证明：

$$\begin{aligned}E[A_1^{2,x}\mid R_x]&=R_x^{-q}\int w\left(\frac{x_1-x}{R_x}\right)(x_1-x)'g^{(2)}(x_1)(x_1-x)f(x_1)\mathrm{d}x_1\\&=R_x^2 f(x)\int w(v)v'g^{(2)}(x)v\mathrm{d}v+(\mathrm{s.o.})\\&=\kappa_2R_x^2f(x)\mathrm{tr}[g^{(2)}(x)]+(\mathrm{s.o.})\end{aligned}$$

因此，由引理 14.1 得

$$\begin{aligned}E[A_1^{2,x}]&=E\{E[A_1^{2,x}\mid R_x]\}\\&=\kappa_2 f(x)\mathrm{tr}\{g^{(2)}(x)\}(k/n)^{2/q}+(\mathrm{s.o.})\\&=(k/n)^{2/q}[f(x)\mu_L(x)+o(1)]\end{aligned}$$

与引理 14.3(ii)的证明类似，我们可以证明

$$E\{[(k/n)^{-2/q}A_1^{2,x}-f(x)\mu_L(x)]^2\}=o(1)$$

因此，$(k/n)^{-2/q}A_1^{2,x}=f(x)\mu_L(x)+o_p(1)$。　$\square$

(ii)的证明:定义$\mathcal{W}_{i,x}=w_{i,x}(X_i-x)(X_i-x)'g^{(2)}(x)(X_i-x)$,我们有

$$
\begin{aligned}
&E[\| A_2^{2,x} \|^2 \mid R_x] \\
&= E[(A_2^{2,x})'(A_2^{2,x}) \mid R_x] \\
&= (2nR_x)^{-2}\sum_i \sum_j E\{[\mathcal{W}_{i,x}]'[\mathcal{W}_{i,x}] \mid R_x\} \\
&= (4nR_x^2)^{-1}E\{[\mathcal{W}_{1,x}'\mathcal{W}_{1,x} \mid R_x]\} + \frac{n-1}{4nR_x{}^2}E\{E[\mathcal{W}_{1,x}' \mid R_x]E[\mathcal{W}_{2,x} \mid R_x]\} \\
&\equiv C_{1n} + C_{2n}
\end{aligned}
$$

运用式(14.33)和引理14.1,容易证明

$$C_{1n} = O_p((k/n)^{2/q}k^{-1}) = o_p((k/n)^{4/q})$$

且$C_{2n}=O((k/n)^{4/q})$(见习题14.2)。因此,我们知道$E[\| A_2^{2,x} \|^2]=E\{E[\| A_2^{2,x} \|^2 | R_x]\}=C_{1n}+C_{2n}=O_p((k/n)^{4/q})$,它意味着$A_2^{2,x}=O_p((k/n)^{2/q})$。

□

439

引理 14.5

(i) $\operatorname{var}(k^{1/2}A_1^{3,x})=c_0\kappa f^2(x)\sigma^2(x)+o(1)$

(ii) $\operatorname{var}(d_nA_2^{3,x})=c_0\kappa_{22}f^2(x)\sigma^2(x)I_q+o(1)$

(iii) $\operatorname{cov}(k^{1/2}A_1^{3,x},d_nA_2^{3,x})=O((k/n)^{1/q})$

(i)的证明:

$$
\begin{aligned}
E[(A_1^{3,x})^2 \mid R_x] &= R_x^{-2q}E\left[w^2\left(\frac{X_1-x}{R_x}\right)u_1^2 \mid R_x\right] \\
&= n^{-1}R_x^{-2q}\int w^2\left(\frac{x_1-x}{R_x}\right)\sigma^2(x_1)f(x_1)\mathrm{d}x_1 \\
&= n^{-1}R_x^{-q}[\kappa\sigma^2(x)f(x) + (\text{s.o.})]
\end{aligned}
$$

因此,由引理14.1得

$$\operatorname{var}(A_1^{3,x}) = E\{E[(A_1^{3,x})^2 \mid R_x]\} = (1/k)\kappa f^2(x)\sigma^2(x) + o(1/k)$$

(ii)的证明:

$$
\begin{aligned}
&E[(A_2^{3,x})(A_2^{3,x})' \mid R_x] \\
&= \frac{1}{nR_x^{2q+4}} \times E\left[w^2\left(\frac{X_i-x}{R_x}\right) \times \sigma^2(X_i)(X_i-x)(X_i-x)' \mid R_x\right] \\
&= (nR_x^{q+2})^{-1}[\kappa_{22}f(x)\sigma^2(x)I_q + o(1)]
\end{aligned}
$$

因此,由引理14.1得

$$\begin{aligned}\text{var}(d_n A_2^{3,x}) &= d_n^2 E\{E[A_2^{3,x}(A_2^{3,x})' \mid R_x]\}\\ &= d_n^2 n^{-1}(k/n)^{(q+2)/q}[\kappa_{22} f(x)\sigma^2(x) I_q + o(1)]\\ &= [\kappa_{22} f(x)\sigma^2(x) I_q + o(1)]\end{aligned}$$ □

440

(iii)的证明：

$$\begin{aligned}&\text{cov}(k^{1/2}A_1^{3,x}, d_n A_2^{3,x} \mid R_x)\\ &= k^{1/2} d_n E[A_1^{3,x} A_2^{3,x} \mid R_x]\\ &= k^{1/2} d_n (nR_x^{2q+2})^{-1} \times \int w^2\left(\frac{x_1 - x}{R_x}\right) \times (x_1 - x)\sigma^2(x_1) f(x_1)\mathrm{d}x_1\\ &= k^{1/2} d_n\ (nR_x^q)^{-1} \times \int w^2(v) vv' \mathrm{d}v m^{(1)}(x) + (\text{s. o.})\end{aligned}$$

其中 $m(x) = \sigma^2(x) f(x)$。

因此，由引理 14.1 我们有

$$\begin{aligned}\text{cov}(k^{1/2}A_1^{3,x}, d_n A_2^{3,x}) &= Ck^{1/2} d_n n^{-1} E(R_x^{-q}) + (\text{s. o.})\\ &= k^{1/2} d_n n^{-1} O(n/k) + (\text{s. o.})\\ &= O((k/n)^{1/q}) = o(1)\\ &\quad (\text{因为 } d_n = k^{1/2}(k/n)^{1/q})\end{aligned}$$

定理 14.5 由引理 14.3 至 14.5 而得。 □

14.9.3 定理 14.10 的证明

引理 14.6 令 $E_1(\cdot) = E(\cdot \mid Z_1 = z_1, R_1)$，且令 $g(z)$ 和 $h(x,z)$ 属于$\mathcal{G}_\nu^\alpha$，对于 $\alpha > 0$，整数 $\nu \geqslant 2$，则

(i) $$\left|E_1\left\{(g(Z_2) - g(z_1)) W\left(\frac{z_1 - Z_2}{R_1}\right)\right\}\right| \leqslant D_g(z_1)\left(\frac{R_1^{q+\nu}}{G(R_1)}\right)$$

(ii) $$\left|E_1\left\{h(X_2, Z_2) w^s\left(\frac{z_1 - Z_2}{R_1}\right)\right\}\right|$$

$$\leqslant L_h(z_1)\left(\frac{R_1^q}{G(R_1)}\right) + M_h(z_1)\left(\frac{R_1^{q+1}}{G(R_1)}\right), s = 1, 2$$

其中 $Z_1 \neq z_1$ 是$(k-1)$-nn 样本中的一个，它落在区间$\{z: \| z - z_1 \| < R_1\}$内，而且当 $D_g(\cdot)$、$L_h(\cdot)$、$M_h(\cdot)$和 $H_h(\cdot,\cdot)$都有有限的 α 阶矩时，对所有$(x, z + u)$和(x,z)有$|h(x,z+u) - h(x,z)| \leqslant H_h(x,z)|u|$。

证明：

(i) 注意条件于(X_1, Z_1, R_1)，Z_2 的密度为$f(z)/G(R_1)$。

441

因此，由 Robinson(1988)中的引理 5，

$$E_1\left\{(g(Z_2)-g(z_1))w\left(\frac{z_1-Z_2}{R_1}\right)\right\}$$

$$=\int[g(z_1+R_1v)-g(z_1)]w(v)f(z_1+R_1v)G(R_1)^{-1}R_1^q\mathrm{d}v$$

$$\leqslant D_g(z_1)\left(\frac{R_1^{q+\nu}}{G(R_1)}\right)$$

(ii)

$$E_1\left\{h(X_2,Z_2)w^s\left(\frac{z_1-Z_2}{R_1}\right)\right\}$$

$$=\frac{R^q}{G(R_1)}\iint h(x,z_1-vR_1)w^s(v)f(x,z+R_1v)\mathrm{d}v\mathrm{d}x$$

$$\leqslant C_1\frac{R_1^q}{G(R_1)}\iint[|h(x,z_1)|+|H_h(x,z_1)vR_1|]w^s(v)\mathrm{d}v\mathrm{d}x$$

$$\leqslant L_h(z_1)\left(\frac{R_1^q}{G(R_1)}\right)+M_h(z_1)\left(\frac{R_1^{q+1}}{G(R_1)}\right)$$

其中 $L_h(z)=C\int|h(x,z)|\mathrm{d}x, M_h(z)=C\int|H_h(x,z)|\mathrm{d}x$。

引理 14.7

$$E[S_{g-\hat{g}}]=O((k/n)^{2/q}k^{-1}b^{-2}+(k/n)^{2\nu/q}b^{-2}=o(n^{-1/2}))$$

其中 $g(z)=\theta(z)$ 或者 $E(X|z)$。

证明:见习题 14.3。 □

引理 14.8

$$S_{g-\hat{g},\varepsilon}=O_p((nk)^{-1/2}(k/n)^{1/q}b^{-1}+n^{-1/2}(k/n)^{\nu/q}b^{-1})$$

其中 $\varepsilon=u$ 或者 $\varepsilon=v\overset{\text{def}}{=}X-E(X|Z)$。

证明:我们只证明当 $g=\theta$ 和 $\varepsilon=u$ 时的情况。当它们为向量时,由柯西不等式可证。

$$E\{S^2_{\hat{g}-g,\varepsilon}\}=n^{-2}\sum_i E[(\hat{g}_i-g_i)^2\varepsilon_i^2I_i]=n^{-1}E[(\hat{g}_1-g_1)^2\varepsilon_1^2I_1]$$

442 $E[(\hat{g}_1-g_1)^2\sigma^2(X_1,Z_1)I_1]$ 和 $E[(\hat{g}_1-g_1)^2I_1]$ 有相同的阶,根据引理 14.7,我们得到

$$n^{-1}E[(\hat{g}_1-g_1)^2I_1]=O((nk)^{-1}(k/n)^{2/q}b^{-2}+n^{-1}(k/n)^{2\nu/q}b^{-2})$$

从而引理 14.8 待证。 □

引理 14.9

(i) $S_{u,\hat{v}}=O_p((nk)^{-1/2}b^{-1})$

(ii) $S_{v,\hat{u}}=O_p((nk)^{-1/2}b^{-1})$

证明：我们只证明(i)中仅当 $v=v(x,z)$ 是一个标量的情况。当 $v=v(x,z)$ 为向量时，该证明可由柯西不等式得到。

(i) $E[S_{u,\hat{v}}^2]=n^{-1}E[u_1^2\hat{v}_1^2 I_1]=n^{-1}E[\sigma^2(X_1,Z_1)\hat{v}_1^2 I_1]$，与习题 14.4 中 $n^{-1}E[\hat{v}_1^2 I_1]=O((nk)^{-1}b^{-2})$ 有相同的阶，接着是引理 14.9。

(ii)(ii)的证明类似于(i)(通过 u 和 v 的相互交换)；因此我们只提供对(i)的证明过程。 □

引理 14.10

(i) $S_{\hat{\varepsilon}}=O_p(k^{-1}b^{-2})$($\varepsilon=u$ 或者 v)

(ii) $S_{\hat{u},\hat{v}}=O_p(k^{-1}b^{-2})$

(i)的证明：

443

$$\begin{aligned}
E[|S_{\hat{\varepsilon}}|] &= E[\hat{\varepsilon}_1^2 I_1] \\
&= n^{-2}E\left\{\left[R_1^{-q}\sum_{j\neq 1}\varepsilon_j w_{1j}I_1/\hat{f}_1\right]^2\right\} \\
&\leqslant n^{-2}b^{-2}E\left[R_1^{-2q}\sum_{j\neq 1}\varepsilon_j^2 w_{1j}^2 I_1\right] \\
&\leqslant C_1(k-1)n^{-2}b^{-2}E\left[R_1^{-2q}w^2\left(\frac{Z_1-\hat{Z}_1}{R_1}\right)\right] \\
&\leqslant Ckn^{-2}b^{-2}E\{E[(R_1^{-q}G(R_1))^{-1}\mid R_1]\} \\
&= kn^{-2}b^{-2}O((n/k)^2) \\
&= O(k^{-1}b^{-2})
\end{aligned}$$

上述过程由引理 14.1 和引理 14.6(ii)可得。

(ii)的证明由(i)和柯西不等式得到。 □

引理 14.11　$S_{\hat{g}-g,\hat{\varepsilon}}=O_p(k^{-1}(k/n)^{1/q}b^{-2}+k^{-1/2}(k/n)^{\nu/q}b^{-2})$，$\varepsilon=u$ 或 v

证明：该引理可由引理 14.7、引理 14.10(i)和柯西不等式得到。 □

引理 14.12　$P[\hat{f}_1<b]=E(1-I_1)=o(1)$

证明：参见习题 14.5。 □

引理 14.13

(i) $S_v=n^{-1}\sum_i v_i v_i'+o_p(1)$

(ii) $S_{v,u}=n^{-1}\sum_i v_i u_i+o_p(n^{-1/2})$

证明：由柯西不等式和引理 14.12 可证。 □

与我们证明定理 7.1 所讨论的类似，引理 14.3 至 14.6 的证明暗含着定理 14.10。

444 14.10 习 题

习题 14.1 证明 $\operatorname{var}(A_{21}^{1,x}) = O(n^{-1})$。

提示:应用引理 14.1。

习题 14.2 证明 $E[A_2^{2,x}] = O_p((k/n)^{2/q})$。

习题 14.3 证明引理 14.7。

提示:证明 $E[S_{g-\hat{g}}] = O((k/n)^{2/q}k^{-1}b^{-2} + (k/n)^{2\nu/q}b^{-2}) = o(n^{-1/2})$。

习题 14.4 证明 $E[\hat{v}_1^2 I_1] = O(k^{-1}b^{-2})$。

习题 14.5 证明引理 14.12。

提示:这是 Liu 和 Lu(1997)的引理 8。

第15章　非参数序列方法

形成这本书主干的核方法只不过是目前已经发展起来的各种非参数技术的一种。在这一章中,我们考虑另一种非参数方法,它通常被称为“序列方法”(series methods)。序列方法根据它们所使用的逼近函数如样条或幂级数很容易被识别出来。序列方法要求使用者选择“节点”(knots)的数量或者幂级数项的数量,在这里,这些节点或者幂级数项的数目类似于核方法中的窗宽的大小。特别是,随着节点数(或者幂级数项数)的增加,得到的拟合变得更加灵活(即更小的逐点偏误但是更大的变异性(方差)),这相当于让与一个核估计量相联系的窗宽缩小。　445

使用序列方法相对核方法来说的一个优势在于可以很容易在最终产生的估计上施加结构(即约束)(见15.3.1节)。我们把序列方法看成是非参数方法的补充,它特别适用于某些类型的问题,如有约束的非参数模型的估计和检验,例如我们在第9章研究的可加和变系数结构的非参数模型。在本章我们将只考虑这样的序列估计量,即它们是序列基本函数的线性组合。对一般序列的估计的深入处理,见Chen(2006)。

15.1　回归函数的估计

446

考虑以下非参数回归模型:

$$Y_i = g(X_i) + u_i, \quad i = 1, \cdots, n \tag{15.1}$$

其中 $g(x) = E(Y_i | x)$ 是(未设定的)条件期望函数。

为便于阐述,我们首先考虑 X_i 为一个标量时的情况。假定独立同分布数据 $\{X_i, Y_i\}_{i=1}^n$,我们可以使用以

$\{p_s(x)\}_{s=1}^K$表示的序列逼近函数估计 $g(\cdot)$，其中 $K=K_n$。逼近函数应具有的性质是，随着 K 的增大，基础函数 $\{p_s(x)\}_{s=1}^K$[①] 的一个线性组合，比如 $\sum_{s=1}^K \beta_s p_s(x)$，从均方误差的意义上来说可以任意好地逼近任何平滑函数 $g(\cdot)$，即存在 $\beta_1,\cdots,\beta_k$ 随着 $K\to\infty$ 使得

$$\mathbf{E}\left\{\left[\sum_{s=1}^K \beta_s p_s(X_i)-g(X_i)\right]^2\right\}\to 0$$

事实上，我们通常需要施加更强的条件，如 $\sup_{x\in\mathcal{S}}\left|\sum_{s=1}^K \beta_s p_s(x)-g(x)\right|=O(K^{-\alpha})$，某些 $\alpha>0$，其中$\mathcal{S}$是一个紧集。这通常需要未知函数 $g(\cdot)$直到一定的阶数都可微。我们已知许多序列函数满足必要的逼近性质。例如，幂级数函数 $\{x^s\}_{s=0}^\infty(p_s(x)=x^s)$，三角函数和样条函数都是完全可以接受的。一个样条函数是分段多项式函数，并且分段线性函数和分段三次多项式(样条)函数都是流行的选择。当使用一个幂级数时，需要为序列中的最高次值(highest term)选择 K 值；当使用一个样条函数时，需要选择节点数，比如，总共 K 个节点。我们下面证明$g(\cdot)$的一致估计要求当 $n\to\infty$ 时，$K\to\infty$，$K/n\to\infty$。这里 K/n 的作用类似于核方法里的平滑参数 h。

最简单的序列基础函数是幂级数，$\{1,x,x^2,\cdots\}$。然而，基于幂级数的估计对一些异常值是敏感的，出于这个原因，幂级数不常用于非参数序列估计。相反，局部分段多项式样条是最流行的选择。一个 r 阶单变量 B 样条基本函数可以给出为(见 Chui(1992，第 4 章)：

447

$$B_r(x\mid t_0,\cdots,t_r)=\frac{1}{a^{r-1}(r-1)!}\sum_{j=0}^r(-1)^j\begin{bmatrix}r\\ j\end{bmatrix}[\max(0,x-t_j)]^{r-1} \tag{15.2}$$

其中 $t_0,\cdots,t_r$ 是在 X 支撑集上的均匀间隔设计节点。当 $r=2$ 时，式(15.2)给出了一个分段线性样条，$r=4$ 时给出了一个分段三次样条(即三阶多项式)。

考虑 X 的取值在单位区间[0,1]的情况。假定我们决定将 X 的支撑分为五个相等间隔($m=5$ 是在 X 支撑里的子区间数目)的区间(有 6 个节点)，因此节点之间的距离将是 $a=1/m=1/5=0.2$。对于 $r=2$(分段线性样条)，我们可以有 4 个函数完全位于单位区间内。但是，对 $x=0$ 以及 $x=1$ 时，这 4 个基础函数的线性组合无法生成非零函数值。为了放宽这一限制，我们在支撑每侧的外面添加 2 个节点，这样总共有 8 个节点。这导致总共 8 个节点具有 6 个基础函

① 基础函数可能依赖于 k，且应被写成$\{p_{sk}(x_i)\}_{s=1}^K$；然而，为了表述简单，下面我们省略了下标 k。

数，即 $K=6=r+m-1(r=2,m=5)$。图 15.1 画出了这 6 个基础函数。如果我们选择 $r=4$(三次样条)，因为每个基础函数将取 4 个子区间 。我们将只有两个函数完全位于单位区间内。我们在单位区间每侧都扩展 3 个额外的节点，导致总共有 12 个节点，此处考虑了 8 个基础函数，即 $K=8=r+m-1(r=4,m=5)$，它们在图 15.2 中被画出。注意当 $n\to\infty$ 时，区间数 m 将会趋于无穷，并且因为 r 是固定的，基础函数的数目也会趋于无穷($K=r+m-1$)。

由图 15.1 和图 15.2 可以看到，基础函数对于所有的 x 值总和为 1。

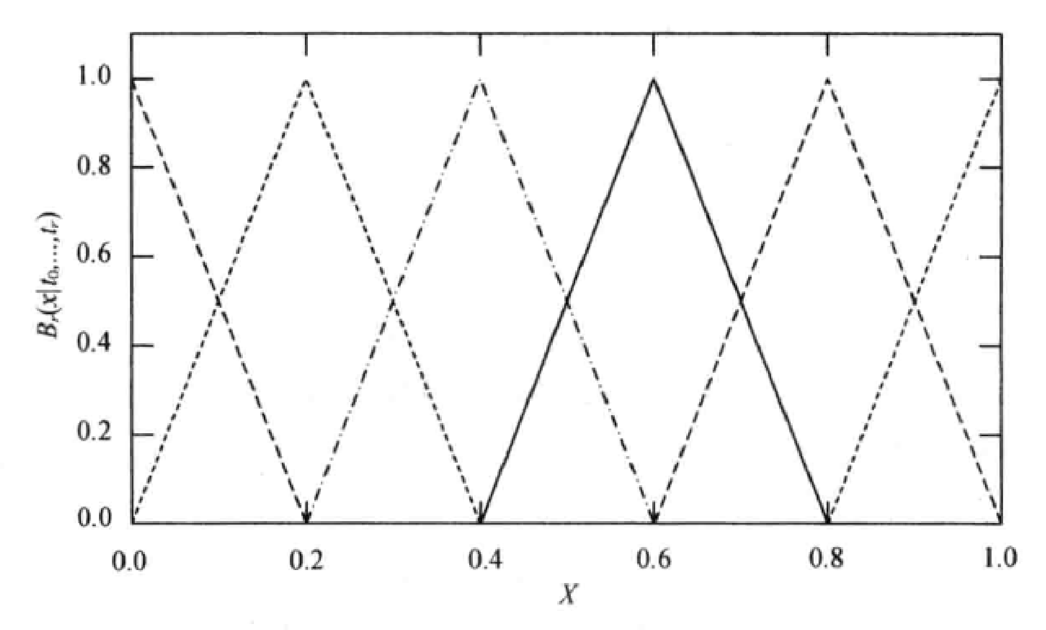


图 15.1 一个 r 阶 B 样条基础函数，有 k 个子区间，$X\in[0,1]$，$r=2,K=5$

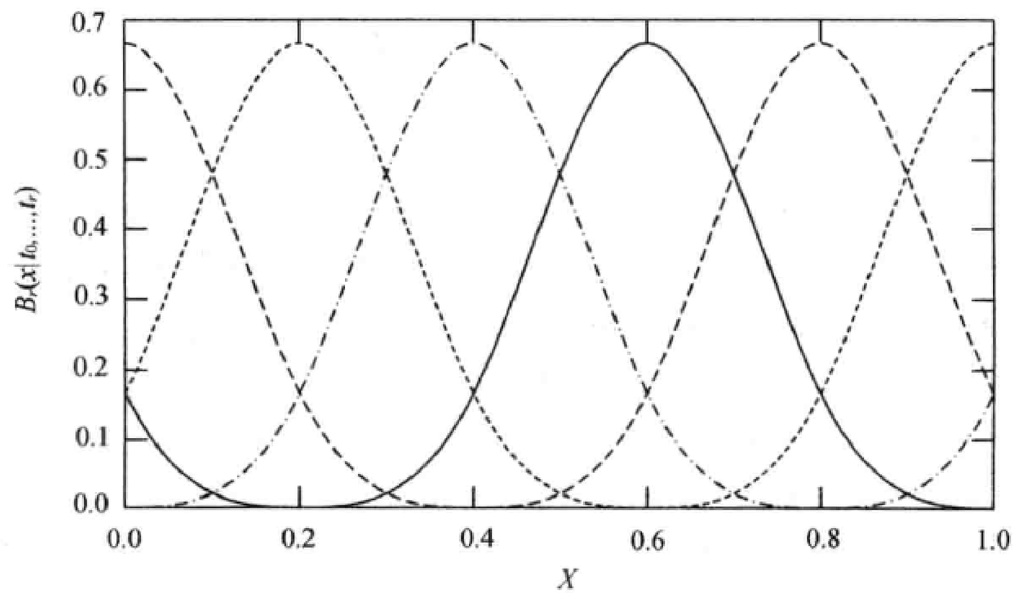


图 15.2 一个 r 阶 B 样条基础函数，有 k 个子区间，$X\in[0,1]$，$r=4,K=5$

对于一般多元回归模型，$X\in\mathbb{R}^q$，可以使用多元序列函数来逼近 $g(x)$。例如，我们可以使用一元样条函数(比如张量样条)去构建多元基础函数。当

$q=2$ 时,幂级数基础函数变为$\{1,x_1,x_2,x_1^2,x_1x_2,x_2^2,\cdots\}$。对于样条函数,如果$(x_1,x_2)$的域(domain)是有界的,我们可以变换这些数据使得(x_1,x_2)都落在一个矩形内,比如$(x_1,x_2)\in[a,b]\times[c,d]$,然后可以使用一元样条函数的乘积去构建多元样条基础函数。如果这个域不能转换为一个矩形,可以使用一个一般多元样条基础函数逼近 $g(x)$;参见 Chui(1988)和 Eubank(1999)关于一般多元样条函数的构建。

448

我们现在讨论怎样用序列方法来估计 $g(x)$。令 $p^K(\cdot)$表示前 K 个逼近函数的 $K\times 1$ 向量(假定 $X\in\mathbb{R}^q$ 是一个连续变量),

449

$$p^K(x) = (p_1(x),\cdots,p_K(x))' \tag{15.3}$$

令 P 表示 $n\times K$ 阶矩阵,其第 i 行由 $p^K(X_i)'$给出,即

$$P = [p^K(X_1),\cdots,p^K(X_n)]'$$

式(15.1)可以重写为

$$Y_i = p^K(X_i)'\beta + \varepsilon_i \tag{15.4}$$

其中 $\varepsilon_i=g(X_i)-p^K(X_i)'\beta+u_i$。根据基础函数的逼近性质,我们可知存在一个 $K\times 1$ 向量 β 使得对所有 x,$p^K(x)'\beta-g(x)\to 0$。我们用 Y 关于 P 的最小二乘回归法估计 β,即

$$\hat{\beta} = (P'P)^{-}P'Y \tag{15.5}$$

其中$(\cdot)^{-}$表示$(\cdot)$的广义逆,Y 是一个 $n\times 1$ 向量,其第 i 行用 Y_i 表示,我们用下式估计 $g(x)$:

$$\hat{g}(x) = p^K(x)'\hat{\beta} \tag{15.6}$$

15.1.1 收敛速度

一些早期的关于序列估计量(如基于样条的估计)的收敛速率的结论可见 Stone(1985,1986),而较近的关于序列估计量的研究包括 Andrews(1991a),Stone(1994),Kooperberg,Stone 和 Truong(1995),Newey(1995,1997),Huang(1998a,1998b,2001),以及 de Jong(2002)。

Andrews(1991a)对关于可能非同分布数据的非参数序列估计量提出了一般的收敛速率和渐近正态性结果。在这一节我们只处理独立同分布数据并且我们使用 Newey(1997)的方法。为了建立 $\hat{g}(x)$的一致性,我们采用下面的来自 Newey 的假定。

假定 15.1 $\{X_i,Y_i\}$是独立同分布数据,与$\{X,Y\}$相同,$\mathrm{var}(Y|x)$在 X 的紧支撑集$\mathcal{S}$上是有界的。

假定 15.2 对于每个 K,存在一个非奇异常数矩阵 B,使得 $P^K(x)=$

$Bp^K(x)$。

(i) 对于每个 K, $E[P^K(X_i)P^K(X_i)']$的最小特征值是远离零有界的。

(ii) 存在一个常数数列$\zeta_0(K)$满足条件 $\sup_{x\in\mathcal{S}}\|P^K(x)\|\leqslant\zeta_0(K)$,其中 $K=$ 450
$K(n)$使得当 $n\to\infty$ 时,$\zeta_0(K)^2K/n\to0$。

(iii) 存在 $\alpha>0$ 使得 $\sup_{x\in\mathcal{S}}|g(x)-p^K(x)'\beta_K|=O(K^{-\alpha})$。

(iv) 当 $n\to\infty$ 时,$K\to\infty$ 且 $K/n\to0$。

以下定理建立了 $\hat{g}(x)$的一致性。

定理 15.1 在假定 15.1 和 15.2 下,可得

(i) $\int[\hat{g}(x)-g(x)]^2\mathrm{d}F(x)=O_p(K/n+K^{-2\alpha})$。

(ii) $n^{-1}\sum_{i=1}^{n}[\hat{g}(X_i)-g(X_i)]^2=O_p(K/n+K^{-2\alpha})$。

(iii) $\sup_{x\in\mathcal{S}}|\hat{g}(x)-g(x)|=O_p(\zeta_0(K)(\sqrt{K}/\sqrt{n}+K^{-\alpha}))$,其中 $F(\cdot)$是 X 的累积分布函数。

定理 15.1 已经被 Newey(1997)证明。因此我们在 15.6 节只提供了对(i)的证明。在 15.6 节我们证明 $O_p(K/n)$项对应于方差项,并且 $K^{-2\alpha}$也对应于偏误平方项。对于样条和幂级数序列,方差项并不直接依赖于 q,而偏误项却直接依赖于 q,因为我们可以证明如果 $g(x)$是 m 次可微的,那么对于样条和幂级数序列,$\alpha=m/q$ 。这与核方法不同但是与 k-nn 方法类似(因为在核估计中,偏误项不依赖于 q,但是方差项是依赖于 q 的)。于是,在定理 15.1 中 $\alpha=m/q$。定理 15.1(i)和(ii)证明幂级数和样条序列的最优速度是 $O_p(n^{-2m/(q+2m)})$,这等于 Stone(1982)的速度。

Newey(1997)给出了使得假定 15.1 和 15.2 成立的关于幂级数和回归样条(B 样条)的原始条件。

假定 15.3

(i) X 的支撑集是紧连接区间的笛卡尔乘积,在此区间 X 有一个绝对连续概率的概率密度函数,它是以一个正常数上有界并远离零有界的。

(ii) $g(x)$在 X 的支撑集上是 m 阶连续可微的。 451

假定 15.4 X 的支撑集是$[-1,1]^q$。

在 X 的支撑集已知且满足假定 15.3(i)的情况下, X 总是可以被重新缩放使得假定 15.4 成立。

Newey(1997, p. 167)证明对于幂级数,假定 15.3(i)意味着对所有 $K(P^K(x)=Bp^K(x)$,$E[P^K(X_i)P^K(X_i)']$的最小特征值是有界的;参见假定 15.2,以及$\zeta_0(K)=O(K)$。同样,接着我们可以由假定 15.3(ii)和 Lorentz(1966)得

知在 $\alpha = m/q$ 下假定 15.2(iii) 成立。于是，假定 15.3 给出了隐含于假定 15.1 和 15.2 的关于幂级数的原始条件。Newey(1997) 也证明了假定 15.3 和 15.4 意味着对于 B 样条 $\zeta_0(K) = O(K)$，假定 15.1 和 15.2 成立。我们在下面的推论中总结这些结果。

推论 15.1

(i) 对于幂级数，如果假定 15.1 和 15.2 满足并且当 $n\to\infty$ 时，$K^3/n\to 0$，那么，如果用 K 代替 $\zeta_0(K)$，则定理 15.1 成立。

(ii) 对于 B 样条，如果假定 15.1、15.3 和 15.4 都满足并且当 $n\to\infty$ 时，$K^2/n\to 0$，那么，当用 $\sqrt{K}$ 代替 $\zeta_0(K)$ 时，定理 15.1 成立。

推论 15.1 表明定理 15.1 的结论在关于幂级数和样条的原始条件下成立。

15.2 序列项 K 的选择

Li(1987) 研究了下面三个著名的关于选择 K 的过程（也可见 Wahba (1985)）。

(i) Mallows 的 C_L（或者 C_p）(Mallows(1973))：选择 $\hat{K}$ 以最小化

$$\hat{K}_C = \arg\min_K n^{-1}\sum_{i=1}^{n}[Y_i - \hat{g}(X_i)]^2 + 2\sigma^2(K/n) \tag{15.7}$$

其中 σ^2 是 u_i 的方差。在实践中，我们可以通过 $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\hat{u}_i^2, \hat{u}_i = Y_i - \hat{g}(X_i)$，来估计 σ^2。

452

(ii) 广义交错鉴定方法(Craven 和 Wahba(1979))：选择 $\hat{K}$ 以最小化

$$\hat{K}_{GCV} = \arg\min_K \frac{n^{-1}\sum_{i=1}^{n}[Y_i - \hat{g}(X_i)]^2}{(1-(K/n))^2} \tag{15.8}$$

(iii) 去一交错鉴定方法(Stone(1974))：选择 $\hat{K}$ 以最小化

$$CV_K = \sum_{i=1}^{n}[Y_i - \hat{g}_{-i}(X_i)]^2 \tag{15.9}$$

其中 $\hat{g}_{-i}(X_i) = p^k(X_i)'\hat{\beta}_{-i}, \hat{\beta}_{-i} = [P'_{-i}P_{-i}]^{-1}P'_{-i}Y_{-i}$ 可以通过移去 (X_i, Y_i) 得到。

注意 GCV 和 C_L 方法在计算上都比去一交错鉴定方法简单。

令 $\hat{G}_K$ 表示 $n\times 1$ 向量，其第 i 个元素是 $\hat{g}(X_i)$。接着我们知道 $\hat{G}_K = P(P'P)^{-1}P'Y \equiv M_n Y$，其中 $M_n = P(P'P)^{-1}P'$ 是 $n\times n$ 投射矩阵。事实上，在 C_L

和 GCV 目标函数的最初的形式中，K 是 $\mathrm{tr}[M_n]$。然而，因为 M_n 是一个投射矩阵，我们有

$$\mathrm{tr}[M_n]=\mathrm{tr}[P(P'P)^{-1}P']=\mathrm{tr}[P'P(P'P)^{-1}]=\mathrm{tr}(I_K)=K$$

Li(1987)证明前面的三种过程都是渐近等价的，并且这三种方法从下式意义上来说都可以得到最优的平滑。

$$\frac{\int[\hat{g}_{\hat{K}}(x)-g(x)]^2\mathrm{d}F(x)}{\inf_K\int[\hat{g}_K(x)-g(x)]^2\mathrm{d}F(x)} \tag{15.10}$$

其中 $\hat{g}_K(x)=\hat{g}(x)$ 正如式(15.6)中所定义的，对于一类 K 成立，并且 $\hat{g}_{\hat{K}}(x)$ 是使用上面所说的方法中的一种选择 K 所得到的 $g(x)$ 的序列估计值，即在式(15.10)中，$\hat{K}=\hat{K}_C$，或者 $\hat{K}=\hat{K}_{\mathrm{GCV}}$，或者 $\hat{K}=\hat{K}_{\mathrm{CV}}$。

等式(15.10)说明在渐近加权积分平方误差等于可能的最小加权积分平方误差的意义上，上述三种方法中的任意一种都可以得到对于 K 的最优选择。

当 K 是最优的选择时，序列估计量的渐近正态结果是很难去建立的。这是因为偏误首项很难计算出来。我们知道 $\mathrm{bias}(\hat{g}(x))=O(K^{-\alpha})$（$\alpha=m/q$，$g(\cdot)$ 是 m 次可微的），但是对于偏误首项，比如说对 $\mathrm{bias}(\hat{g}(x))=B(x)K^{-\alpha}+o(K^{-\alpha})$ 中的 $B(x)$，是很难得到一个准确的结果的。Zhou，Shen 和 Wolfe(1998)在某些严格条件下，如节点是渐近等间隔的，以及样条的自由度 m 等于 $p-1$，对属于 C^p（即对于回归函数来说有连续的 p 阶导数）的一元样条回归函数得到一个渐近偏误首项表达式。关于一般情形下使用样条的渐近偏误的困难更加详细的讨论见 Huang(2003)。因此，当 K 是最优选择时，一般的时间序列估计值的渐近正态结果是得不到的。然而，如果引入一个平滑不足条件，我们很容易就可以得到一个时间序列估计值的渐近正态结果。

453

15.2.1　渐近正态性

给定 $\hat{g}(x)$ 的偏误首项表达式不可得，为了得到 $\hat{g}(x)$ 的渐近分布，我们要引入平滑不足条件，即我们假定 $K^{-2\alpha}=o(K/n)$。于是，偏误平方比方差项有更小的阶，使得 $\sqrt{n/K}[\hat{g}(x)-g(x)]$ 有一个渐近的零均值。

假定 15.5

(i) $E\{[y-g(x)]^4|x\}$ 是有界的，并且 $\mathrm{var}(y|x)$ 远离零有界。

(ii) $K^{-\alpha}=o((K/n)^{1/2})$。

给定平滑不足条件，由引理 15.2 的 $(P'P/n)=Q+o_p(1)\equiv I+o_p(1)$，容易看出 $\hat{g}(x)$ 的渐近方差是 $n^{-1}V_K$，其中

$$V_K=p^K(x)'Q^{-1}\Sigma Q^{-1}p^K(x)=p^K(x)'\Sigma p^K(x)$$

因为

$$Q = E(p^K(X_i)p^K(X_i)') = I_K$$

$$\Sigma = E[p^K(X_i)p^K(X_i)'\sigma^2(X_i)]$$

且 $\sigma^2(X_i) = E(u_i^2 \mid X_i)$。注意 $V_K = O(K)$，比如，如果误差是条件同方差的，即 $E(u_i^2 \mid X_i) = E(u_i^2) = \sigma^2$，那么 $V_K = \sigma^2 p^K(x)'p^K(x) = O(K)$。因此，$\text{var}(\hat{g}(x)) = O(K/n)$。

定理 15.2 在假定 15.1、15.2 和 15.5 下

$$\sqrt{n}\hat{V}_K^{-1/2}(\hat{g}(x) - g(x)) \xrightarrow{d} N(0,1)$$

454 其中

$$\hat{V}_K = p^K(x)'\hat{Q}^{-1}\hat{\Sigma}\hat{Q}^{-1}p^K(x)$$

$$\hat{Q} = n^{-1}\sum_{i=1}^{n} p^K(X_i)p^K(X_i)'$$

$$\hat{\Sigma} = n^{-1}\sum_{i=1}^{n} p^K(X_i)p^K(X_i)'\hat{u}_i^2$$

$$\hat{u}_i = Y_i - p^K(X_i)'\hat{\beta}$$

定理 15.2 是 Newey(1997)定理 2 的一个特殊情形，其中也包括了一个 $\hat{g}$ 的一般线性函数的情形。关于条件同方差误差情形，定理 15.2 的证明留作习题(见习题 15.1)。

15.3 一个部分线性模型

Donald 和 Newey(1994)考虑使用序列方法估计一个以下形式的部分线性模型

$$Y_i = Z_i'\gamma + g(X_i) + u_i \tag{15.11}$$

其中 Z_i 是 γ 维的，X_i 是 q 维的。

序列方法涉及对于某些 $K \times 1$ 向量 β 用 $p^K(x) = (p_1(x), p_2(x), \cdots, p_K(x))'$ 的线性组合，比如 $p^K(x)'\beta$，逼近未知函数 $g(x)$。重写式(15.11)可得

$$Y_i = Z_i'\gamma + p^K(X_i)'\beta + \varepsilon_i \tag{15.12}$$

其中 $\varepsilon_i = g(X_i) - p^K(X_i)'\beta + u_i$，我们用 Y_i 关于 $(Z'_i, p^K(X_i)')$ 的最小二乘回归法估计 γ 和 β，即

$$\hat{\xi} = \begin{bmatrix}\hat{\gamma}\\ \hat{\beta}\end{bmatrix} = (\mathcal{W}'\mathcal{W})^{-1}\mathcal{W}'y, \quad \mathcal{W} = (Z, P) \tag{15.13}$$

其中 Z 是 $n \times \gamma$ 维的，P 是 $n \times K$ 维的。

使用分块逆(partitioned inverse),我们可以证明(见习题 15.2):

$$\hat{\gamma} = [(Z-\hat{Z})'(Z-\hat{Z})]^{-1}(Z-\hat{Z})'(Y-\hat{Y}) \tag{15.14}$$

其中 $\hat{Z}$是一个 $n\times\gamma$ 向量,其第 i 行是 $\hat{Z}_i = p^K(X_i)'(P'P)^{-1}P'Z$,即 $E(Z_i|X_i)$ 的序列估计值,而 $\hat{Y}$是一个 $n\times 1$ 向量,其第 i 个元素由 $\hat{Y}_i = p^K(X_i)'(P'P)^{-1}P'Y$ 给出,即 $E(Y_i|X_i)$的序列估计值。

455

等式(15.14)与在第 7 章中给出的半参数核(或者 k-nn)估计值有同样的形式。Donald 和 Newey(1994)证明了下面的结果。

定理 15.3　在假定 15.1 和 15.2 下,并且假定$(K/n+K^{-2\alpha})=o(n^{-1/2})$,那么$\sqrt{n}(\hat{\gamma}-\gamma)\to N(0,V)$依分布收敛,其中 V 与第 7 章定理 7.1 中的定义相同。

对定理 15.3 的证明在 15.6 节给出。

除了假定 15.1 和 15.2 外,定理 15.3 还要求附加的假定:$K/n+K^{-2\alpha}=o(n^{-1/2})$,即非参数序列估计均方误差有比 $n^{-1/2}$更小的阶。这和我们在第 7 章中引入的半参数核方法所用的条件是类似的,在那里我们也是假定非参数估计均方误差比 $n^{-1/2}$更小。

关于 $g(x)$的非参数估计值由下式给出

$$\hat{g}(x) = p^K(x)'\hat{\beta} \tag{15.15}$$

可以证明 $\hat{\beta}=(P'P)^{-1}P'(Y-Z\hat{\gamma})$。注意 $\hat{\gamma}-\gamma=O_p(n^{-1/2})$, $\hat{g}(x)$的渐近分布是由 $\tilde{g}(x)=p^K(X_i)'\hat{\beta}$ 决定的,其中 $\tilde{\beta}=(P'P)^{-1}P'(Y-Z\hat{\gamma})$。显然,$\tilde{g}(x)$与定理 15.1 有同样的渐近分布。于是,我们知道当 $\tilde{g}(x)$ 如式(15.15)中定义时,定理 15.1 成立。也就是说,用 $F(\cdot)$表示 X 的累积分布函数,那么

(i) $\int[\hat{g}(x)-g(x)]^2\mathrm{d}F(x) = O_p(K/n+K^{-2\alpha})$

(ii) $n^{-1}\sum_{i=1}^{n}[\hat{g}(X_i)-g(X_i)]^2 = O_p(K/n+K^{-2\alpha})$

(iii) $\sup_{x\in\mathcal{S}}|\hat{g}(x)-g(x)| = O_p(\zeta_0(K)(\sqrt{K}/\sqrt{n}+K^{-\alpha}))$

15.3.1　一个可加的部分线性模型

在第 9 章我们讨论了使用核方法的可加部分线性模型的估计。在这一节我们考虑用序列方法来估计这些模型。与使用需要边际积分的核方法相比,使用序列方法估计可加的部分线性模型拥有很多优势。核方法涉及首先估计一个高维的非参数模型,然后使用边际积分的方法得到一个潜在函数的低维估计量。然而,这种方法可能导致有限样本的效率损失,因为可加结构这个事实并没有在最初的估计过程中用到。相反,序列方法在整个估计过程中可以很容易

456

地引入可加结构。

考虑下面的一个可加的部分线性模型：

$$Y_i = Z_i'\gamma + g_1(X_{1i}) + g_2(X_{2i}) + \cdots + gL(X_{Li}) + u_i \qquad (15.16)$$

其中 Z_i 是包含一个常数项的随机变量的 $r\times 1$ 向量，$\gamma=(\gamma_1,\cdots,\gamma_r)'$ 是一个关于未知参数的 $r\times 1$ 向量，并且 X_{li} 维度为 $q_l(q_l\geqslant 1, l=1,\cdots,L)$。用 X_i 表示从 $(X_{1i},\cdots,X_{Li})$ 得到的非重叠变量。X_i 的维度是 q 且 $L\leqslant q\leqslant\sum_{l=1}^{L}q_l$。$E(u_i|X_i,Z_i)=0$ 且 $g_1(\cdot),\cdots,g_L(\cdot)$ 都是未知的平滑函数。

单独的函数 $g_l(\cdot)(l=1,\cdots,L)$ 在没有施加一些识别条件时不能被识别。对于核方法，一个方便的识别条件是对所有 $l=2,\cdots,L$，施加 $E[g_l(X_{li})]=0$。对序列方法的识别条件可以通过选择一些在序列逼近基础函数中很容易引入的规范化规则来得到。比如，在一个没有交互作用项的可加模型中，

$$g(x_1,\cdots,x_q) = c + g_1(x_1) + \cdots + g_q(x_q)$$

其中 $x_j\in\mathbb{R}$，我们可以使用 $g_j(x_j=0)\equiv g_j(0)=0$ 作为识别条件。

类似地，对于一个具有二阶交互项的可加模型情形，我们有

$$g(x_1,\cdots,x_q) = c + \sum_{j=1}^{q}m_j(x_j) + \sum_{j=1}^{q-1}\sum_{l>j}^{q}m_{jl}(x_j,x_l) \qquad (15.17)$$

式(15.17)的表达形式不是唯一的，但是我们可以通过施加下面的识别条件来做到：

$$m_j(x_j=0)\equiv m_j(0)=0,\quad j=1,\cdots,q$$
$$m_{jl}(0,x_l)=m_{jl}(x_j,0)=0,\quad \text{所有 } x_j,x_l \text{ 值}(1\leqslant j<l\leqslant q) \qquad (15.18)$$

原则上我们总是可以在逼近基础函数上施加识别条件。令 $\mathcal{N}_1=\{1,2,\cdots\}$ 表示正整数集合。如果我们使用 $\{\phi_t(x_j)\}_{t\in\mathcal{N}_1}$，$j=1,\cdots,q$ 的一个有限线性组合作为基础函数去逼近可加函数 $m_j(x_j)$，上面的识别条件意味着我们应该使用 $\{\phi_t(x_j)\}_{t\in\mathcal{N}_1}$ 的一个有限线性组合去逼近 $m_j(x_j)$，其中对所有 $t\in\mathcal{N}_1$，有 $\phi_t(x_j=0)=\phi_t(0)=0$。然后我们可以使用 $\{\phi_t(x_j)\phi_s(x_l)\}_{s>t\in\mathcal{N}_1}$ 的一个有限线性组合逼近 $m_{jl}(x_j,x_l)$。比如，考虑一种多项式（幂级数）序列有 $\phi_t(x_j)=x_j^t$ 的情形。关于 $m_j(x_j)$ 的逼近函数是 $\{x_j^t\}_{t\in\mathcal{N}_1}=\{x_j,x_j^2,x_j^3,\cdots\}$（没有常数项）的一个有限线性组合，使得 $\phi_t(0)=0^t=0$ 满足。并且逼近 $m_{jl}(x_j,x_l)$ 的基础函数是 $\{x_j^t x_l^s\}_{t,s\in\mathcal{N}_1}=\{x_jx_l, x_j^2x_l, x_jx_l^2, x_j^2x_l^2,\cdots\}$ 的一个有限线性组合。由于通过式(15.18)施加条件，逼近函数具有性质 $\phi_t(x_j=0)\phi_s(x_l)=\phi_t(x_j)\phi_s(x_l=0)=0$。可以直接对有更高阶的交互项的可加函数将上面的识别条件一般化。

现在我们讨论 γ 的识别条件。对于任何随机变量（向量）$\mathcal{A}_i$，用 $E_{\mathcal{G}}(\mathcal{A}_i)$ 表

示$\mathcal{A}_i$在线性可加函数空间$\mathcal{G}$上的投射。在所有的可加函数$\mathcal{G}$中，$E_{\mathcal{G}}(\mathcal{A}_i)$是$\mathcal{A}_i$的最接近的函数（从均方误差的意义上来说）。

注释 15.1　令$\mathcal{V}_i=\mathcal{A}_i-E_{\mathcal{G}}(\mathcal{A}_i)$，那么$E_{\mathcal{G}}(\mathcal{V}_i)=0$。也就是说，对任意的随机变量（向量）$\mathcal{A}_i$，我们可以正交分解$\mathcal{A}_i=E_{\mathcal{G}}(\mathcal{A}_i)+\mathcal{V}_i$，其中$E_{\mathcal{G}}(\mathcal{A}_i)\in\mathcal{G}$且$\mathcal{V}_i\perp\mathcal{G}$，即$E_{\mathcal{G}}(\mathcal{V})=0$。

下面我们给出一类可加函数的定义。

定义 15.1　如果(i) $\xi(x)=\sum_{l=1}^{L}\xi_l(x_l)$，$\xi_l(x_l)$在其支撑集$\mathcal{S}_l$上是连续的，其中$\mathcal{S}_l$是$\mathbb{R}^{q_l}(l=1,\cdots,L)$的一个紧子集。(ii) $\sum_{l=1}^{L}E[\xi_l(x_l)]^2<\infty$，(iii) $\xi_l(0)=0,l=2,\cdots,L$，我们说一个函数$\xi(x)$属于一个可加类函数$\mathcal{G}_A(\xi\in\mathcal{G}_A)$。

当$\xi(x)$是一个向量值函数时，如果ξ的每个分量都属于$\mathcal{G}_A$，我们说$\xi\in\mathcal{G}_A$。用向量矩阵表示法，我们可以把式(15.16)写为

$$\mathcal{Y}=\mathcal{Z}\gamma+g_1+g_2+\cdots+g_L+\mathcal{U}\equiv\mathcal{Z}\gamma+g+U \tag{15.19}$$

其中$\mathcal{Y}$和U都是$n\times 1$向量，其第i个分量由Y_i和u_i分别得出，$\mathcal{Z}$是$n\times r$向量，其第i个分量由Z_i'给出，并且g是$n\times 1$向量，其第i个分量由$g_i=g(X_i)\equiv\sum_{l=1}^{L}g_l(X_{li})$给出。

458

我们使用K_l函数的一个线性组合去逼近$g_l(x_l)(l=1,\cdots,L)$。

$$p_l^{K_l}(x_l)=(p_{l1}^{K_l}(x_l),\cdots,p_{lK_l}^{K_l}(x_l))'$$

因此，我们使用$K=\sum_{l=1}^{L}K_l$函数的一个线性组合$(p_1^{K_1}(x_1)',\cdots,p_L^{K_L}(x_L)')\equiv p^K(z)'$去逼近$g(x)=g(x_1,\cdots,x_l)=\sum_{l=1}^{L}g_l(x_l)$。逼近函数$p^K(x)$具有以下性质：

(i) $p^K(x)\in\mathcal{G}_A$;

(ii) 当K_l增加时（对所有$l=1,\cdots,L$），$p^K(x)$的一个线性组合在均方误差意义上可以任意好地逼近任何$g\in\mathcal{G}_A$。

我们引入一些记号。定义

$$p_l=(p_l^{K_l}(x_{l1}),\cdots,p_l^{K_l}(x_{ln}))'(l=1,\cdots,L),\quad P=(p_1,\cdots,p_L) \tag{15.20}$$

注意p_l是$n\times K_l$维的，且P是$n\times K$维的。

令$M=P(P'P)^-P'$，其中$(\cdot)^-$表示$(\cdot)$的任一对称广义逆。对于一个$n\times 1$或者$n\times r$矩阵A，定义$\tilde{A}=MA$。接着用M左乘式(15.19)可得

$$\tilde{\mathcal{Y}} = \tilde{\mathcal{Z}}\gamma + \tilde{g} + \tilde{U} \tag{15.21}$$

用式(15.19)减式(15.21)可得

$$\mathcal{Y} - \tilde{\mathcal{Y}} = (\mathcal{Z} - \tilde{\mathcal{Z}})\gamma + g - \tilde{g} + U - \tilde{U} \tag{15.22}$$

我们通过$\mathcal{Y} - \tilde{\mathcal{Y}}$关于$\mathcal{Z} - \tilde{\mathcal{Z}}$的最小二乘回归估计$\gamma$,即

$$\hat{\gamma} = [(\mathcal{Z} - \tilde{\mathcal{Z}})'(\mathcal{Z} - \tilde{\mathcal{Z}})]^{-}(\mathcal{Z} - \tilde{\mathcal{Z}})'(\mathcal{Y} - \tilde{\mathcal{Y}}) \tag{15.23}$$

然后用$\hat{g}(x) = p^K(x)'\hat{\beta}$估计$g(x) = \sum_{l=1}^{L} g_l(x_l)$,其中$\hat{\beta}$被给出为

$$\hat{\beta} = (P'P)^{-}P'(\mathcal{Y} - \tilde{\mathcal{Z}}\hat{\gamma}) \tag{15.24}$$

在下面给出的条件下,$(P'P)$和$(\mathcal{Z} - \tilde{\mathcal{Z}})'(\mathcal{Z} - \tilde{\mathcal{Z}})$都是渐近非奇异的。于是,当我们取极限$\min\{K_1,\cdots,K_n\}\to\infty$时(当$n\to\infty$时),所有的广义逆事实上都是可逆的。注意当$(\mathcal{Z} - \tilde{\mathcal{Z}})'(\mathcal{Z} - \tilde{\mathcal{Z}})$和$(P'P)$都是非奇异时,在式(15.23)和式(15.24)中给出的$\hat{\gamma}$和$\hat{\beta}$在数值上等于由$\mathcal{Y}$关于$(\mathcal{Z},P)$做回归时得到的最小二乘估计值。在有限样本的应用中,$(\mathcal{Z} - \tilde{\mathcal{Z}})'(\mathcal{Z} - \tilde{\mathcal{Z}})$和/或$(P'P)$可能是奇异的。然而,可以去掉多余的回归元使得这些矩阵非奇异。

459

接下来需要建立关于$\hat{\gamma}$的渐近分布和$\hat{g}(x) = p^K(x)'\hat{\beta}$到$g(x)$的收敛速度的假定。

假定 15.6

(i) 与(Y,X,Z)一样,$(Y_1,X_1,Z_1),\cdots,(Y_n,X_n,Z_n)$是独立同分布的,$(X,Z)$的支撑集是$\mathbb{R}^{r+q}$上的紧子集。

(ii) 定义$\theta(x) = E(Z|X=x)$。$\theta(x)$和$\text{var}(Y|X=x,Z=z)$在(X,Z)的支撑集上都是有界函数。

假定 15.7

(i) 当$f=g$或者$f=\theta_s$时(其中$s=1,\cdots,r$且$\theta(x) = E(Z_i|x) = (\theta_1(x)),\cdots,\theta_r(x))'$),存在某些$\delta_l > 0$ $(l=1,\cdots,L)$, $\beta_f = \beta_{fK} = (\beta'_{fK_1},\cdots,\beta'_{fK_L})'$,使得当$\min\{K_1,\cdots,K_L\}\to\infty$时,

$$\sup_{x\in\mathcal{Z}} | f(x) - P^K(x)'\beta_f = O\left(\sum_{l=1}^{L} K_l^{-\delta_l}\right)$$

(ii) 当$n\to\infty$时,$\sqrt{n}(\sum_{l=1}^{L} K^{-\delta_l})\to 0$。

假定 15.6 和 15.7(i)在估计可加函数模型的文献中是标准的。假定

15.7(ii)要求数据是平滑不足的,因此偏误估计有比 $n^{-1/2}$ 更小的阶。

令 $E_A(Z_i)$ 表示在可加函数空间$\mathcal{G}_A$内 Z_i 的投射(在 L_2 范数下)。也就是说,$E_A(Z_i)$是一个属于$\mathcal{G}_A$(它具有一个可加的结构)的元素并且它是在均方误差意义上在$\mathcal{G}_A$内的所有随机函数中最接近 Z_i 的。接下来的定理给出了 $\hat{\gamma}$ 的渐近分布。

定理 15.4 定义 $\varepsilon_i = Z_i - E_A(Z_i)$,并且假定 $\Phi \stackrel{\text{def}}{=} E[\varepsilon_i \varepsilon_i']$是正定的,那么在假定 15.1、15.2、15.6 和 15.7 下,我们有

(i) $\sqrt{n}(\hat{\gamma} - \gamma) \to N(0, \Sigma)$ 依分布收敛,其中

$$\Sigma = \Phi^{-1} \Omega \Phi^{-1}, \quad \Omega = E[\sigma_u^2(X_i, Z_i) \varepsilon_i \varepsilon_i']$$

$$\sigma_u^2(x, z) = E(u_i^2 \mid X_i = x, Z_i = z)$$

460

(ii) Σ 的一个一致估计量可以由 $\hat{\Sigma} = \hat{\Phi}^{-1} \hat{\Omega} \hat{\Phi}^{-1}$ 给出,其中

$$\hat{\Phi} = n^{-1} \sum_i (Z_i - \tilde{Z}_i)(Z_i - \tilde{Z}_i)', \quad \hat{\Omega} = n^{-1} \sum_i \hat{u}_i^2 (Z_i - \tilde{Z}_i)(Z_i - \tilde{Z}_i)'$$

$\tilde{Z}_i'$ 是 $\tilde{\mathcal{Z}} = M \mathcal{Z} (M = P(P'P)^- P')$的第 i 行,且 $\hat{u}_i = Y_i - Z_i' \hat{\gamma} - \hat{g}(X_i)$ 。

定理 15.4 的证明在 Li(2000)中给出。

$\Phi = E(\varepsilon_i \varepsilon_i')$是正定的是关于 γ 的一个识别条件。正如第 9 章所讨论的,这个条件比当忽略$\sum_{l=1}^{L} g_l(x_l)$的可加结构时识别 γ 需要的条件要弱一些。

回忆 $v_i = Z_i - E(Z_i \mid X_i)$ 以及定义 $\eta_i = E(Z_i \mid X_i) - E_A(Z_i)$,则 $\varepsilon_i = v_i + \eta_i$。使用 $E(v_i \mid X_i) = 0$,我们得到

$$\Phi = E[(v_i + \eta_i)(v_i + \eta_i)'] = E(v_i v_i') + E(\eta_i \eta_i')$$

于是,$E(v_i v_i')$是正定的或者 $E(\eta_i \eta_i')$是正定的都意味着 Φ 是正定的。注意当估计 γ 时如果我们忽略$\sum_{l=1}^{L} g_l(x_l)$的可加结构,$E(v_i v_i')$是一个正定矩阵是必要的识别条件(对于 γ)。于是,通过使用模型有一个可加的部分线性结构的信息,我们可以把对 γ 的识别条件弱化为 Φ 是一个正定矩阵。

当误差是条件同方差时,即 $E(u_i^2 \mid X_i, Z_i) = E(u_i^2) = \sigma_u^2$。那么,在$\sqrt{n}(\hat{\gamma} - \gamma)$的渐近方差的逆等于半参数有效边界的意义上,$\hat{\gamma}$ 是半参数有效的。

有意思的是,我们观察到当 $E(X \mid z)$不是一个可加函数时,即使 X 是一个关于 Z(即 $E(Z \mid X) = Z$)的确定性函数,γ 仍可识别。考虑一个简单的情形 $L = 2$,其中 X_{1i}和 X_{2i}都是标量。令 $Z_i = X_{1i} X_{2i}$,于是模型(15.16)变成

$$Y_i = Z_i \gamma + g_1(X_{1i}) + g_2(X_{2i}) + u_i \tag{15.25}$$

式(15.25)中的参数 γ 是可被识别的,因为 $Z_i = X_{1i} X_{2i}$不是一个关于 X_{1i}和

X_{2i}的可加可分离的函数。模型(15.25)的优势在于它只涉及一维的非参数回归函数,因此它避免了维数诅咒。它也允许一个交互项(以参数形式进入模型)并且因此它比不允许交互项的可加函数更加具有一般性。在实践中,我们可以以任何另一个已知的(非可加的)(X_{1i},X_{2i})的函数代替交互项 $Z_i=X_{1i}X_{2i}$。显然如果我们忽略可加结构 $g(x)=g_1(x_1)+g_2(x_2)$,则 γ 在式(15.25)中是不能被识别的。

461

下面的定理给出了 $\hat{g}(x)=p^K(x)'\hat{\beta}$ 到 $g(x)=\sum_{l=1}^{L}g_l(x_l)$ 的收敛速度。

定理 15.5 在假定 15.6 和 15.7 下,令$\mathcal{S}$表示 z 的支撑,我们有

(i) $\sup_{x\in\mathcal{S}}|\hat{g}(x)-g(x)|=O_p(\zeta_0(K))\left[\sqrt{K}/\sqrt{n}+\sum_{l=1}^{L}K_l^{-\delta_l}\right]$;

(ii) $n^{-1}\sum_{i=1}^{n}[\hat{g}(X_i)-g(X_i)]^2=O_p\left(K/n+\sum_{l=1}^{L}K_l^{-2\delta_l}\right)$;

(iii) $\int[\hat{g}(x)-g(x)]^2\mathrm{d}F(x)=O_p\left(K/n+\sum_{l=1}^{L}K_l^{-2\delta_l}\right)$,其中 $F(\cdot)$是 Z 的累积分布函数。

定理 15.5 基本上说明 $\hat{g}(x)$到 $g(x)$的收敛速度是相同的,无论 γ 是否已知或者是否在构建 $\hat{g}(x)$的时候使用估计值 $\hat{\gamma}$。这是可以预期的,因为 $\hat{\gamma}-\gamma=O_p(n^{-1/2})$,它比非参数(序列)估计量比如 $\hat{g}(x)$的收敛速度要快。

我们也可以通过 $\hat{g}_l(x_l)=p_l^{K_l}(x_l)'\hat{\beta}_l$ 来估计 $g_l(x_l)$,其中 $\hat{\beta}_l$ 是一个 $K_l\times 1$ 向量,它可以从 $\hat{\beta}=(\hat{\beta}_1',\cdots,\hat{\beta}_L')'$得到,并且$\mathcal{S}_l$表示它的支撑。

当 x_l 都是标量并且每一个未知函数$(g_l(\cdot))$都是 c 阶可微时,我们可以选择所有具有相同阶数的 K_l,即对所有 $l=1,\cdots,L,K_l=K/L$,并且对于幂级数关于 K 的条件变成 $K^3/n\to 0$ 以及 $nK^{-2c}\to 0$(要求 $c>3/2$),而对于样条 $K^2/n\to 0$ 和 $nK^{-2c}\to 0$,要求 $c>1$。

15.3.2 非线性可加成分的选择

考虑一个非独立数据的可加模型如下:

$$Y_t=\beta_0+\sum_{j=1}^{d}g_j(X_{tj})+u_t \tag{15.26}$$

其中函数 $g_j(\cdot)$形式是未知的。我们可以再次引入 $g_j(0)=0$ $(j=1,\cdots,d)$作为识别条件。我们假定只有变量$\{X_{tj},j\in s\}$,$s\in\{1,\cdots,d\}$的一个子集,有尽可能小的基数 $q\leqslant d$,提供了(几乎)与 $X_t=(X_{t1},\cdots,X_{td})$相同的关于 Y_t 的信息,也就是说

$$E(Y_t \mid x_{tj}=x_j, j\in s)=E(Y_t \mid X_t=x),\text{几乎所有 } x \tag{15.27}$$

被选择的变量被称为“显著性变量”。如果 X_t 只包含 Y_t 的滞后值，被选择 462 的滞后期被称为“显著性滞后”。Huang 和 Yang(2004)建议使用样条(即非参数序列方法)去估计未知函数(即 $g_j(\cdot)$)。为了明确，我们写成

$$g_j(X_{tj})\cong\sum_{l=1}^{L_j}\gamma_{lj}B_{lj}(X_{tj}) \tag{15.28}$$

其中 $B_{lj}(\cdot)(l=1,\cdots,L_j)$ 是对一个给定自由度和节点序列，满足识别条件 $B_{lj}(0)=0$ 的 B 样条基础函数。我们通过最小二乘法估计 β_0 和 γ 系数，也就是最小化准则函数

$$\sum_{t=1}^{n}\left[Y_t-\beta_0\sum_{j=1}^{d}\sum_{l=1}^{L_j}\gamma_{lj}B_{lj}(X_{tj})\right]^2 \tag{15.29}$$

令 $\hat{\beta}_0$ 和 $\hat{\gamma}_{lj}$表示最小化式(15.29)的值，$g(x_t)$ 的基于样条的估计值由下式给出：

$$\hat{g}(X_t)=\hat{\beta}_0+\sum_{j=1}^{d}\sum_{l=1}^{L_j}\hat{\gamma}_{lj}B_{lj}(X_{tj}) \tag{15.30}$$

令 MSE_s 表示在集合 s 内使用变量的子集对均方误差的估计，即

$$\mathrm{MSE}_s=\frac{1}{n}\sum_{t=1}^{n}\left[Y_t-\tilde{\beta}_0-\sum_{j\in s}\sum_{l=1}^{L_j}\tilde{\gamma}_{tj}B_{lj}(X_tj)\right]^2 \tag{15.31}$$

其中 $\tilde{\beta}_0$ 和 $\tilde{\gamma}_{lj}$都是 β_0 与 γ_{lj}的最小二乘估计量，并且对 $j\in s, l=1,\cdots,L_j,\gamma_{lj}$的估计量只使用了基础函数 $B_{lj}(X_{lj})$ 的子集。

Huang 和 Yang(2004)建议使用 BIC 准则选择显著性变量的集合，即选择最小化下式的变量 s 的集合：

$$\mathrm{BIC}_s=\ln(\mathrm{MSE}_s)+\frac{N_s}{n}\ln(n) \tag{15.32}$$

其中 $N_s=1+\sum_{j\in s}L_j$ 是在序列估计量中的(在变量 s 的集合内)待估计的参数数量。

因为技术上的原因，Huang 和 Yang(2004)把以上的估计问题限制在一个 463 紧集内。令 $\hat{s}$表示通过 BIC 准则选择出的变量集合，并且令 s_0 表示显著性变量的真集，Huang 和 Yang 证明 BIC 选择准则是一致的，也即

$$\lim_{n\to\infty}\mathrm{P}[\hat{s}=s_0]=1 \tag{15.33}$$

也就是说，当样本容量增加至无穷时，由 BIC 方法选择出正确的变量集合的概率收敛于 1。

当然，式(15.33)依赖于显著性变量集合 s_0 是被唯一定义的。Huang 和 Yang(2004)证明给定 $X_t=(X_{t1},\cdots,X_{td})$ 的联合密度函数在 X_t 上是连续的，s_0 确

实是被唯一被定义的。

Huang 和 Yang(2004)报告的模拟证明 BIC 准则对于选择正确的显著性变量集合运行很好。Huang 和 Yang(2004)也采用这种方法选择显著性滞后期,其使用的是美国季度失业率数据(1948 年第四季度到 2003 年第一季度),对数据取了四阶差分从而消除非平稳性。令 $Y_t = r_{t+4} - r_t$,那么总样本大小是 $n = 217$,$\{Y_t\}_{t=1}^{217}$。为了得到从 208 期到 217 期滚动的样本外预测值,他们舍弃了最后 10 个观测值。BIC 方法选择了 2 个显著性滞后期,其均方预测误差(MSPE)是 0.023,然而一个同样根据 BIC 选择的具有 5 个滞后期的线性模型给出的 MSPE 是 0.058。于是,具有 2 个滞后期的非参数可加模型可以得到一个比具有 5 个滞后期的线性模型小 50% 的均方预测误差,暗示了在美国失业率数据中显著的非线性。

Härdle 等(2001)提出一种可供选择的(基于小波的)方法去检验是否某些可加部分是多项式结构的(比如线性)而不用设定剩余部分的结构,这也解决了选择显著性协变量这个问题。

15.3.3 用一个已知的联接函数估计可加模型

Horowitz 和 Mammen(2004)考虑用一个已知的联接函数(link function)来估计可加函数:

$$Y_i = G[c_0 + g_1(Z_{1i}) + \cdots + g_q(Z_{qi})] + u_i, \quad i = 1, \cdots, n \quad (15.34)$$

464

其中 $G(\cdot)$ 是一个已知的联接函数,g_α 是未知的单变量函数。我们讨论如何在第 9 章的边际积分法的基础上去估计式(15.34)。边际积分法的一个问题在于在最初的估计过程中必须估计一个 q 维函数 $g(z) = E(Y_i | Z_i = z)$。

序列估计方法可以通过使用一个可加基础函数很容易地引入可加结构。然而,序列方法的一个缺点是,当序列项的项数是最优选择时,它的渐近分布一般是未知的,因为准确的偏误首项通常都是未知的。Horowitz 和 Mammen(2004)的方法结合了序列和核方法的优势。他们建议在第一步使用序列方法使得可以很容易引入可加结构。然后,在第二步中,可以使用核方法去估计单独的函数 g_α。

假定 z 的支撑集是 $[-1,1]^q$,Horowitz 和 Mammen(2004)要求识别条件 $\int_{-1}^{1} g_\alpha(z_\alpha)\mathrm{d}z_\alpha = 0$,$\alpha = 1, \cdots, q$。令 $\{p_l : l = 1, \cdots\}$ 表示平滑函数在 $[-1,1]$ 区间里的一个基础。施加在 p_l 上的条件是 $\int_{-1}^{1} p_l(v)\mathrm{d}v = 0$,以及 delta 函数 $\int_{-1}^{1} p_l(v)p_j(v)\mathrm{d}v = \delta_{lj}$ 等于 1(如果 $l = j$)或者 0(其他),于是,对所有 $\alpha = 1, \cdots,$

$q, z_\alpha \in [-1,1]$以及某些合适的系数 $\theta_{\alpha l}$，可得（基础函数完成了）

$$g_\alpha(z_\alpha) = \sum_{l=1}^{\infty} \theta_{\alpha l} p_l(z_\alpha) \tag{15.35}$$

可以很容易检查 $\theta_{lj} = \int_{-1}^{1} g_\alpha(z_\alpha) p_l(z_\alpha) dz_\alpha$（见习题 15.5）。令 $P^K(z)$表示由下式定义的可加基础函数：

$$P^K(z) = [1, p_1(z_1), \cdots, p_k(z_1), p_1(z_2), \cdots, p_k(z_2), \cdots, p_1(z_q), \cdots, p_k(z_q)]' \tag{15.36}$$

注意 P^K 是$(kq+1) \times 1$ 维的。令 $P^K(z')\theta$ 逼近 $c_0 + \sum_{\alpha=1}^{q} g_\alpha(z_\alpha)$。在第一步我们估计 θ 为

$$\hat{\theta} = \arg\min_{\theta \in \Theta_k} n^{-1} \sum_{i=1}^{n} \{Y_i - G[P^K(Z_i)'\theta]\}^2 \tag{15.37}$$

其中 $\Theta \in \mathbb{R}^{qk+1}$是一个紧参数集。$c_0 + g(z)$的序列估计值是 $P^K(z')\hat{\theta} = \tilde{c}_0 + \tilde{g}(z)$，其中 $\tilde{c}_0$ 是 $\hat{\theta}$ 的第一个元素。

为了得到 $g_\alpha(z_\alpha)$的第二步估计值，令 $Z_{\underline{\alpha} i}$表示去掉 $Z_{\alpha i}$时的 Z_i，定义

$$\tilde{g}_{\underline{\alpha}}(Z_{\underline{\alpha} i}) = \sum_{s \neq \alpha}^{q} \tilde{g}_s(Z_{si})$$

其中

465

$$\tilde{g}_s(Z_{si}) = p^k(Z_{si})'\theta_s, \quad p^k(Z_{si}) = (p_1(Z_{si}), \cdots, p_k(Z_{si}))'$$

且 θ_s 是对应于 $p^k(z_s) = (p_1(z_s), \cdots, p_k(z_s))'$的 $k \times 1$ 系数向量。令 $\omega_h(v) = \omega(v/h)$是一个单变量二阶核函数。令 $G^{(j)}(v) = \partial^j G(v)/\partial v^j$，$j=1,2$，并且对 $j=0,1$，定义

$$\begin{aligned} S_{nj\alpha}^{(1)}(z_\alpha, \tilde{g}) = &-2 \sum_{i=1}^{n} \{Y_i - G[\tilde{c}_0 + \tilde{g}_\alpha(Z_\alpha) + \tilde{g}_{\underline{\alpha}}(Z_{\underline{\alpha i}})]\} \\ &\times G^{(1)}[\tilde{c}_0 + \tilde{g}_\alpha(z_\alpha) + \tilde{g}_{\underline{\alpha}}(Z_{\underline{\alpha i}})](Z_{\alpha i} - z_\alpha)^j w_h(Z_\alpha - Z_{\alpha i}) \end{aligned} \tag{15.38}$$

同时对 $j=0,1,2$，定义

$$\begin{aligned} S_{nj\alpha}^{(2)}(z_\alpha, \tilde{g}) = &2 \sum_{i=1}^{n} - G[\tilde{c}_0 + \tilde{g}_\alpha(z_\alpha) + \tilde{g}_{\underline{\alpha}}(Z_{\underline{\alpha i}})](Z_{\alpha \iota} - z_\alpha)^j w_h(z_\alpha - Z_{\alpha i}) \\ &- 2 \sum_{i=1}^{n} \{Y_i - G[\tilde{c}_0 + \tilde{g}_\alpha(z_\alpha) + \tilde{g}_{\underline{\alpha}}(Z_{\underline{\alpha i}})]\} \\ &\times G^{(2)}[\tilde{c}_0 + \tilde{g}_\alpha(z_\alpha) + \tilde{g}_\alpha(Z_{\underline{\alpha i}})](Z_{\alpha i} - z_\alpha)^j w_h(z_\alpha - Z_{\alpha i}) \end{aligned} \tag{15.39}$$

关于 $g_\alpha(z_\alpha)$ 的第二步估计量是

$$\hat{g}_\alpha(z_\alpha) = \tilde{g}_\alpha(z_\alpha) - \frac{S_{n2\alpha}^{(2)}(z_\alpha,\tilde{g})S_{n0\alpha}^{(1)}(z_\alpha,\tilde{g}) - S_{n1\alpha}^{(2)}(z_\alpha,\tilde{g})S_{n1\alpha}^{(1)}(z_\alpha,\tilde{g})}{S_{n0\alpha}^{(2)}(z_\alpha,\tilde{g})S_{n2\alpha}^{(2)}(z_\alpha,\tilde{g}) - S_{n1\alpha}^{(2)}(z_\alpha,\tilde{g})} \quad (15.40)$$

由式(15.40)给出的估计量有一个很好的直观解释。如果 $\tilde{c}_0$ 和 $g_{\underline{\alpha}}$ 都是 c_0 和 $g_{\underline{\alpha}}$ 的真值，$g_\alpha(z_\alpha)$ 的局部线性估计将会是下面给出 b_0 的值，其中(b_0,b_1)最小化下面的目标函数：

$$S_{n\alpha}(z_\alpha,b_0,b_1) = \sum_{i=1}^{n}\{Y_i - G[\tilde{c}_0 + b_0 - b_1(Z_{\alpha i} - z_\alpha) + \tilde{g}_{\underline{\alpha}}(Z_{\underline{\alpha i}})]\}^2 w_h(z_\alpha - Z_{\alpha i}) \quad (15.41)$$

同时注意 $S_{nj\alpha}^{(1)} = \partial S_{n\alpha}/\partial b_j(j=0,1)$是在 $b_0 = \hat{g}_\alpha(z_\alpha)$ 和 $b_1 = 0$ 处估值的。$S_{nj\alpha}^{(2)}$ 给出了在同一点估值的 $S_{n\alpha}$ 的二阶导数。估计量(15.40)是使用一个 Newton 步骤的结果，采用初始值 $b_0 = \tilde{g}_\alpha(z_\alpha)$，$b_1 = 0$，逼近式(15.41)右边的最小值。为了得到 $\hat{g}_\alpha(z_\alpha)$ 的渐近分布，我们首先定义一些相关的等式。定义

466

$$D_0(z_\alpha) = 2\int G^{(1)}[c_0 + g_\alpha(z_\alpha) + g_{\underline{\alpha}}(z_{\underline{\alpha}})]^2 f_z(z_\alpha,z_{\underline{\alpha}})\mathrm{d}z_{\underline{\alpha}}$$

$$D_1(z_\alpha) = 2\int G^{(1)}[c_0 + g_\alpha(z_\alpha) + g_{\underline{\alpha}}(z_{\underline{\alpha}})]^2[\partial f_z(z_\alpha,z_{\underline{\alpha}})/\partial z_{\underline{\alpha}}]\mathrm{d}z_{\underline{\alpha}}$$

$$m(z) = G^{(2)}[c_0 + g_\alpha(z_\alpha) + g_{\underline{\alpha}}(z_{\underline{\alpha}})]g_\alpha^{(1)}(z_\alpha) + G^{(1)}[c_0 + g_\alpha(z_\alpha) + g_{\underline{\alpha}}(z_{\underline{\alpha}})]g_\alpha^{(2)}(z_\alpha)$$

$$\beta_\alpha(z_\alpha) = 2C_h^2\kappa_2 D_0(z_\alpha)^{-1}\int m(z_\alpha,z_{\underline{\alpha}}) \times G^{(1)}[c_0 + g_\alpha(z_\alpha) + g_{\underline{\alpha}}(z_{\underline{\alpha}})]f_z(z_\alpha,z_{\underline{\alpha}})\mathrm{d}z_{\underline{\alpha}}$$

$$V_\alpha(z_\alpha) = \kappa C_h^2 D_0(z_\alpha)^{-1}\int \mathrm{var}(u_i \mid z)G^{(1)}[g(z)]f_z(z)\mathrm{d}z_{\underline{\alpha}}$$

其中 $\kappa_2 = \int\omega(v)v^2\mathrm{d}v$ 且 $\kappa = \int\omega(v)^2\mathrm{d}v$。

在一些正则条件下，包括 $h_\alpha = c_h n^{-1/5}$，$k = c_k n^{4/15+v}$，对某些常数 $0 < c_h, c_k < \infty$ 以及 $0 < v < 1/30$，并且 G 和 g_α 是二阶连续可微的，Horowitz 和 Mammen(2004)证明了下面的结果：

$$\sqrt{nh_\alpha}[\hat{g}_\alpha(z_\alpha) - g_\alpha(z_\alpha) - h_\alpha^2\mu_\alpha(z_\alpha)] \xrightarrow{d} N(0,V_\alpha(z_\alpha)) \quad (15.42)$$

进一步地，对于 $\alpha \neq s$，$\sqrt{nh_\alpha}[\hat{g}_\alpha(z_\alpha) - g_\alpha(z_\alpha)]$ 以及 $\sqrt{nh_s}[\hat{g}_s(z_s) - g_s(z_s)]$ 都是渐近独立的。

Horowitz 和 Lee(2005)进一步扩展了 Horowitz 和 Mammen(2004)对于估计一个可加的分位数回归模型的结果。文中他们建议首先使用非参数序列方法(用可加基础函数)来估计所有的可加分位数函数。然后,在第二步使用一个局部多项式核方法估计一个单独的分位数函数。他们证明在一些非常普通的条件下这些估计值有相当神奇的性质。

15.4　部分线性变系数模型的估计

在第 7 章我们讨论了一个以下形式的部分线性模型:

$$Y_i = v_i'\gamma + \delta(X_i) + u_i, \quad i = 1, \cdots, n \tag{15.43}$$

其中 $v_i'\gamma$ 是参数成分并且 $\delta(x_i)$ 是一个构成模型非参数成分的未知函数。这个模型可以扩展为半参数变系数模型

467

$$Y_i = v_i'(X_i) + \delta(X_i) + u_i, \quad i = 1, \cdots, n \tag{15.44}$$

其中 $\gamma(x)$ 是关于 X 的未知的平滑函数向量。定义 $X_i = (1, \nu_i')'$ 且令 $\beta(x) = (\delta(x), \gamma(x)')'$,则式(15.44)可以更紧致地写为

$$Y_i = Z_i'\beta(X_i) + u_i, \quad i = 1, \cdots, n \tag{15.45}$$

变系数模型在很多种背景下都是适用的,比如说,在一个截面生产函数框架中,$v_i = (\text{Labor}_i, \text{Capital}_i)'$代表了公司的劳动力和资本投入,并且 $x_i = \text{R\&D}_i$ 是公司的研究和开发支出。变系数模型意味着劳动力和资本投入的系数可能随着公司的研发投入的多少而直接变化,所以劳动和资本的边际生产率依赖于公司的研发水平。另一方面,部分线性模型(15.43)只允许研发变量对生产函数有一个中性的效应,即它只改变生产前沿的水平但是不会影响劳动和/或资本的边际生产率。

然而,当某些 β 系数事实上是常数时,我们应该把它们作为常数来建模,从而通过吸收这个信息得到更有效的估计。再次考虑生产函数的例子。如果把资本进一步分成流动资本和固定资本,我们预期研发水平将会影响固定资本的边际生产率而不是流动资本的边际生产率是合理的。这就给出了一个以下形式的部分线性变系数模型:

$$Y_i = W_i'\gamma + Z_i'\beta(X_i) + u_i, \quad i = 1, \cdots, n \tag{15.46}$$

其中 W_i 是一个随机变量的 $d \times 1$ 向量,γ 是一个 $d \times 1$ 的未知参数向量,z_i 是 $r \times 1$ 维的,$X_i = (X_{i1}, \cdots, X_{iq})$ 是 q 维的,$\beta(\cdot) = (\beta_1(\cdot), \cdots, \beta_r(\cdot))'$是一个未知变系数函数的 $r \times 1$ 维向量,并且 u_i 是误差项,满足 $E(u_i | W_i, X_i, Z_i) = 0$。

我们使用序列方法估计式(15.46)。对 $l = 1, \cdots, r$,我们用 k_l 基础函数的一个线性组合 $p_l^{k_l}(x)'\alpha_l^{k_l}$ 逼近变系数函数 $\beta_l(z)$,其中 $p_l^{k_l}(x) = [p_{l1}(x), \cdots, p_{lk_l}(x)]'$是

一个基础函数的 $k_l \times 1$ 向量，$\alpha_l^{k_l}=(\alpha_{l1},\cdots,\alpha_{lk_l})'$是一个未知参数的 $k_l \times 1$ 向量。

468 定义 $K \times 1$ 矩阵

$$p^K(X_i,Z_i) = (Z_{i1}p_1^{k_1}(X_i)',\cdots,Z_{ir}p_r^{k_r}(X_i)')'$$

且 $\alpha=(\alpha_1^{k_1'},\cdots,\alpha_r^{k_r'})'$，其中 $K=\sum_{l=1}^{r}k_l$。我们使用一个 K 函数的线性组合 $p^K(Z_i,X_i)'\alpha$去逼近 $Z_i'\beta(X_i)$，于是我们把式(15.46)重写为

$$\begin{aligned}Y_i &= W_i'\gamma + p^K(Z_i,X_i)'\alpha + (Z_i'\beta(X_i) - p^K(Z_i,X_i)'\alpha) + u_i\\ &= W_i'\gamma + p_i^K(Z_i,X_i)'\alpha + \text{error}_i \qquad (15.47)\end{aligned}$$

其中 error_i 的定义是显而易见的。

我们现在介绍一些矩阵记号。令$\mathcal{Y}=(Y_1,\cdots,Y_n)'$, $u=(u_1,\cdots,u_n)'$, $\mathcal{W}=(W_1,\cdots,W_n)'$，且

$$P = (p^K(Z_1,X_1),\cdots,p^K(Z_n,X_n))'$$

于是，模型(15.47)可以用向量矩阵表示法重写为

$$\mathcal{Y} = \mathcal{W}\gamma + P_\alpha + \text{error} \qquad (15.48)$$

令 $\hat{\gamma}$和 $\hat{\alpha}$表示根据式(15.48)对$\mathcal{Y}$关于$(\mathcal{W},P)$进行回归得到的 γ 和 α 的最小二乘估计值，我们用 $\hat{\beta}_l(x)\overset{\text{def}}{=}p_l^{k_l}(x)'\hat{\alpha}_l(l=1,\cdots,r)$来估计 $\beta_l(x)$。我们将要建立关于 $\hat{\gamma}$的$\sqrt{n}$正态性结果并且推导 $\hat{\beta}_l(x)$的收敛速度。

我们给出一个 $\hat{\gamma}$和 $\hat{\alpha}$的替代形式，以便于下面给出的渐近分析。式(15.46)可以用矩阵形式重写为

$$\mathcal{Y} = \mathcal{W}\gamma + G + u \qquad (15.49)$$

其中 $G=(X_1'\beta(Z_1),\cdots,X_n'\beta(Z_n))'$。

定义 $M=P(P'P)^-P'$，其中$(\cdot)^-$表示$(\cdot)$的任一对称广义逆。对于 $n \times m$ 矩阵 A，定义 $\tilde{A}=MA$。然后，对式(15.49)两边同时左乘 M，得到

$$\tilde{Y} = \tilde{W}_\gamma + \tilde{G} + \tilde{u} \qquad (15.50)$$

用式(15.49)减去式(15.50)得到

$$\mathcal{Y}-\tilde{\mathcal{Y}} = (\mathcal{W}-\tilde{\mathcal{W}})\gamma + (G-\tilde{G}) + u - \tilde{u} \qquad (15.51)$$

$\hat{\gamma}$也可以通过$\mathcal{Y}-\tilde{\mathcal{Y}}$对$\mathcal{W}-\tilde{\mathcal{W}}$的最小二乘回归得到，即

$$\hat{\gamma} = [(\mathcal{W}-\tilde{\mathcal{W}})'(\mathcal{W}-\tilde{\mathcal{W}})]^-(\mathcal{W}-\tilde{\mathcal{W}})'(\mathcal{Y}-\tilde{\mathcal{Y}}) \qquad (15.52)$$

469 同时，$\hat{\alpha}$ 可以由式(15.48)得到，只需用 $\hat{\gamma}$ 代替 γ，

$$\hat{\alpha} = (P'P) - P'(\mathcal{Y}-\mathcal{W}\hat{\gamma}) \qquad (15.53)$$

由此，我们可得 $\hat{\beta}_l(x)=p_l^{k_l}(x)'\hat{\alpha}_l^{k_l}$, $l=1,\cdots,r$。

在下面给出的假定下，$(\mathcal{W}-\tilde{\mathcal{W}})'(\mathcal{W}-\tilde{\mathcal{W}})$和 $P'P$ 都是渐近非奇异的。因此，由式(15.52)和式(15.53)给出的 $\hat{\gamma}$和 $\hat{\alpha}$ 都是很好定义的，并且在数值上与通过$\mathcal{Y}$对$(\mathcal{W},P)$回归得到的最小二乘估计值相等。

接下来，我们给出一个定义和一些假定用于推导出下面给出的主要结果。

定义 15.2　$g(x,z)$据称属于变系数类函数$\mathcal{G}$，如果(i) 对某些连续性函数 $h_l(x)$有 $g(x,z)=z'h(x)\equiv\sum_{l=1}^{r}z_lh_l(x)$，其中 $h(x)=(h_1(x),\cdots,h_r(x))'$，并且 (ii) $\sum_{l=1}^{r}E[z_{il}^2h_l(x_i)^2]<\infty$，其中 $z_l(z_{il})$是 $z(z_i)$的第 l 个元素。

对于任意函数 $f(x,z)$，令 $E_{\mathcal{G}}[f(x,z)]$表示 $f(x,z)$在变系数函数空间$\mathcal{G}$上的投射(在 L_2 范数下)。也就是说，$E_{\mathcal{G}}[f(x,z)]$是一个属于$\mathcal{G}$的元素，并且它是在$\mathcal{G}$内所有函数中最接近 $f(x,z)$的函数。更具体地说(x_l 是 x 的第 l 个元素，$l=1,\cdots,r$)，

$$E\{(f(x,z)-E_{\mathcal{G}}[f(x,z)])(f(x,z)-E_{\mathcal{G}}[f(x,z)])'\}$$
$$=\inf_{\sum_l z_lh_l(x)\in\mathcal{G}}E\left\{(f(x,z)-\sum_{l=1}^{r}z_lh_l(x))(f(x,z)-\sum_{l=1}^{r}z_lh_l(x))'\right\} \tag{15.54}$$

定义 $\theta(x,z)=E[W|x,z]$，且令 $m(x,z)=E_{\mathcal{G}}[\theta(x,z)]$。接下来的假定用于建立 $\hat{\gamma}$的渐近分布和 $\hat{\beta}(z)$的收敛速率。

假定 15.8

(i) $(Y_i,W_i,X_i,Z_i)_{i=1}^{n}$如$(Y_1,W_1,X_1,Z_1)$是独立同分布的，且$(W_1,X_1,Z_1)$的支撑集是$\mathbb{R}^{d+q+r}$的紧子集。

(ii) $\theta(x_1,z_1)$和 $\mathrm{var}(Y|w_1,x_1,z_1)$都是在$(w_1,x_1,z_1)$的支撑上的有界函数。

470

假定 15.9

(i) 对于$f(x,z)=\sum_{l=1}^{r}z_l\beta_l(x)$或者$f(x,z)=m_j(x,z)$($m_j$ 是 m 的第j 个分量 $(j=1,\cdots,d)$)，存在某些 $\delta_l>0(l=1,\cdots,r)$，$\alpha_f=\alpha_{fK}=(\alpha_1^{k'_1},\cdots,\alpha_r^{k'_r})$，使得

$$\sup_{(x,z)\in\mathcal{S}}|f(x,z)-P^K(x,z)'\alpha_f|=O\left(\sum_{l=1}^{r}k_l^{-\delta_l}\right)$$

(ii) $\min\{k_1,\cdots,k_r\}\to\infty$，当 $n\to\infty$ 时，$\sqrt{n}\left(\sum_{l=1}^{r}k_l^{-2\delta_l}\right)\to 0$。

在上面的假定下，我们现在可以陈述我们的主要定理。

定理 15.6 定义 $\varepsilon_i = \omega_i - m(X_i, Z_i)$，其中 $m(X_i, Z_i) = E_{\mathcal{G}}(W_i)$，并且假定 $\Phi \equiv E[\varepsilon_i \varepsilon_i']$ 是正定的，于是在假定 15.8 和 15.9 下，我们可得

(i) $\sqrt{n}(\hat{\gamma} - \gamma) \to N(0, \Sigma)$ 依分布收敛，其中

$$\Sigma = \Phi^{-1} \Omega \Phi^{-1}, \quad \Omega = E[\sigma^2(W_i, X_i, Z_i) \varepsilon_i \varepsilon_i']$$

且 $\sigma^2(W_i, X_i, Z_i) = E[u_i^2 | W_i, X_i, Z_i]$。

(ii) Σ 的一致估计由 $\hat{\Sigma} = \hat{\Phi}^{-1} \hat{\Omega} \hat{\Phi}^{-1}$ 给出，其中

$$\hat{\Phi} = n^{-1} \sum_{i=1}^{n} (W_i - \tilde{W}_i)(W_i - \tilde{W}_i)', \quad \hat{\Omega} = n^{-1} \sum_{i=1}^{n} \hat{u}_i^2 (W_i - \tilde{W}_i)(W_i - \tilde{W}_i)'$$

W_i 是 $\tilde{\mathcal{W}}$ 的第 i 行并且 $\hat{u}_i = Y_i - W_i' \hat{\gamma} - p^K(X_i, Z_i)' \hat{\alpha}$。

定理 15.6 的证明在 15.6 节给出。

在条件同方差的误差假定下（即 $E[u_i^2 | W_i, X_i, Z_i] = E(u_i^2) = \sigma^2$），在 $\sqrt{n}(\hat{\gamma} - \gamma)$ 的渐近方差的逆等于半参数有效边界的意义上，$\hat{\gamma}$ 的估计值是半参数有效的（见 Chamberlain(1992)）。

接下来的定理给出了对 $l = 1, \cdots, r$，$\hat{\beta}_l(x) = p_l^{k_l}(x)' \hat{\alpha}_l^{k_l}$ 到 $\beta_l(x)$ 的收敛速率。

定理 15.7 在假定 15.8 和 15.9 下，令 $\mathcal{S}_x$ 表示 x_i 的支撑集，然后我们可以得到，对于 $l = 1, \cdots, r$，

(i) $\sup_{x \in \mathcal{S}_x} | \hat{\beta}_l(x) - \beta_l(x) | = O_p\left(\zeta_0(K) \left(\sqrt{K}/\sqrt{n} + \sum_{l=1}^{r} k_l^{-\delta_l} \right) \right)$；

(ii) $\frac{1}{n} \sum_{i=1}^{n} (\hat{\beta}_l(x) - \beta_l(x))^2 = O_p\left(K/n + \sum_{l=1}^{r} k_l^{-2\delta_l} \right)$；

(iii) $\int (\hat{\beta}_l(x) - \beta_l(x))^2 \mathrm{d}F_x(x) = O_p\left(K/n + \sum_{l=1}^{r} k_l^{-2\delta_l} \right)$，其中 F_x 是 x_i 的累积分布函数。

471 定理 15.7 的证明和定理 15.1 的证明是类似的，所以在此省略。

可以直接将变系数模型的渐近正态结果应用于局部线性变系数模型以得到 $\hat{\beta}(z)$ 的渐近分布。这是因为 $\hat{\gamma} - \gamma = O_p(n^{-1/2})$，它收敛到零的速度比任何非参数估计都要快。因此，无论使用 $\hat{\gamma}$ 的估计值或者 γ 的真值，$\hat{\beta}(z)$ 都具有同样的渐近分布，后者成为一个变系数模型（当 γ 未知时），且 Huang 和 Shen(2004) 以及 Huang, Wu 和 Zhou(2004) 的结论是适用的。也可参见 Chen 和 Conley (2001) 对于面板时间序列的半参数空间模型的相关研究。

我们已经证明非参数序列估计方法可用于方便地引入形状约束，比如可加结构和变系数结构。非参数序列方法也可以用于引入其他类型的约束，比如单

调性；参见 He 和 Shi(1998)对非参数估计单调 B 样条的构建。

15.4.1　正确参数回归函数形式的检验

考虑以下原假定，检验是否正确地设定参数回归函数形式，即 H_0^a：$E(Y|x)=m(x,\gamma)$对于几乎所有 x 成立。令 $\hat{\gamma}$表示基于参数原假定模型的 γ 的一个$\sqrt{n}$一致估计量，并且令 $\hat{g}(x)$表示 $g(x)=E(Y|x))$的非参数序列估计值，即 $\hat{g}(x)=p^K(x)'(P'P)^{-1}P'Y$，其中 $P^K(x)$、P 和 Y 与 15.1 节中的定义一样。定义 $\hat{u}_i=Y_i-m(X_i,\hat{\gamma})$并且令 $\hat{v}_i=\hat{g}(X_i)-m(X_i,\hat{\gamma})$。Hong 和 White(1995)建议基于下式检验 H_0^a：

$$\hat{I}_{se}^a=\frac{1}{n}\sum_{i=1}^{n}\hat{v}_i\hat{u}_i \tag{15.55}$$

式(15.55)背后的逻辑是

$$E[u_i(g(X_i)-m(X_i,\gamma))]=E[(g(X)-m(X,\gamma))^2]\geqslant 0$$

当且仅当 H_0^a 为真时等号成立。

假定 15.10　$\{X_i',Y_i\}_{i=1}^n$是独立同分布过程，有 $E(Y_i^4)<\infty$。X_i 在$\mathcal{S}$上具有一个连续的正的概率密度函数 f，此处$\mathcal{S}$是 X_i 的紧支撑集。误差是条件同方差的。

Hong 和 White(1995)证明了下列结论。　472

定理 15.8　在假定 15.2 和 15.10 下，如果 H_0^a 是真的，那么

$$\hat{J}_{se}^a=(n\hat{I}_{se}^a/\hat{\sigma}_{se,a}^2-K)/(2K)^{1/2}\xrightarrow{d}N(0,1) \tag{15.56}$$

其中 $\hat{\sigma}_{se,a}^2=n^{-1}\sum_{i=1}^{n}\hat{u}_i^2$，$\hat{u}_i=Y_i-m(x_i,\hat{\gamma})$。

Hong 和 White(1995)也讨论了怎样扩展他们的检验以处理条件异方差误差的情形。

Hong 和 White(1995)对 H_0^a 提出了另一种基于序列的检验统计量，它基于一个参数回归模型和非参数回归模型之间残差平方和的差，即

$$\tilde{I}_{se}^a=\frac{1}{n}\sum_{i=1}^{n}\hat{\nu}_i\hat{u}_i^2-\frac{1}{n}\sum_{i=1}^{n}\hat{\nu}_i\hat{\eta}_i^2 \tag{15.57}$$

其中 $\hat{\eta}_i=Y_i-\hat{g}(x_i)$是非参数估计量的残差，和定理 15.8 一样，Hong 和 White(1995)证明在 H_0^a 下，

$$\tilde{J}_{se}^a=(n\tilde{I}_{se}^a/\hat{\sigma}_{se,a}^2-K)/(2K)^{1/2}\xrightarrow{d}N(0,1) \tag{15.58}$$

如果 H_0^a 是非真的，$\hat{I}_{se}^a$和 $\tilde{I}_{se}^a$都会依概率收敛于某个正常数；因此，$\hat{J}_{se}^a$和 $\tilde{J}_{se}^a$

将会以速率$\sqrt{n}/\sqrt{K}$发散至$+\infty$,即它们是一致检验。

在第 12 章中我们构造了基于$E[u_i E(u_i|X_i)]$的核检验。我们也可以用这种方法构造基于序列的检验。我们通过$\hat{r}(x)=p^K(x)'[P'P]^{-1}P'\hat{u}$估计$r(x)\equiv E(u|x)$,并且我们用下式定义$r(X_i)$的去一估计值:

$$\hat{r}_{-i}(X_i)=p^K(X_i)'[P'P]^{-1}\sum_{j\neq i}p^K(X_j)\hat{u}_j \tag{15.59}$$

我们可以基于下式给出的$E[u_i E(u_i|X_i)]$的样本对应值构造一个检验统计量

$$\bar{I}^a_{se}=\frac{1}{n}\sum_{i=1}^n\hat{u}_i\hat{r}_{-i}(X_i)=\frac{1}{n}\sum_{i=1}^n\sum_{j\neq i}^n\hat{u}_i p^K(X_i)'(P'P)^{-1}p^K(X_j)\hat{u}_j$$

Sun 和 Li(2006)证明在H_0^a假定下

$$\bar{J}^a_{se}=n\bar{I}^a_{n,se}/[(2K)^{1/2}\hat{\sigma}^2_{se,a}]\xrightarrow{d}N(0,1) \tag{15.60}$$

473 对式(15.60)的证明留作习题(见习题 15.6)。

事实上,我们可以通过下式计算$I^a_{n,se}$:

$$\begin{aligned}\bar{I}^a_{n,se}&=\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n\hat{u}_i p^K(X_i)'[P'P]^{-1}p^K(X_j)\hat{u}_j\\&\quad-\frac{1}{n^2}\sum_{i=1}^n\hat{u}_i^2 p^K(X_i)'[P'P]^{-1}p^K(X_i)\end{aligned}$$

在H_0^a下, Hong 和 White(1995)要求

$$E\{[\theta_0(X_i)-p^K(X_i)'\beta_K]^2\}=o(K^{1/4}n^{1/2})$$

然而这个条件对$\bar{J}^a_{se}$统计量(在H_0^a下)来说是没有要求的。同时,$\bar{J}^a_{se}$没有一个非零的中心值。因此,它可能比拥有非零中心项的检验表现出更好的有限样本性质。Sun 和 Li(2006)报告的模拟结果确实证明在有限样本中$\bar{J}_{n,se}$检验比$\hat{J}_{n,se}$和$\tilde{J}_{n,se}$检验表现得好。这些检验都可以扩展以处理条件异方差误差。为了简洁,我们只给出$\bar{J}_{n,se}$检验的结果。

定理 15.9 在假定 15.2 和 15.10 下,但是允许条件异方差误差,即$E(u_i^2|X_i)=\sigma^2(X_i)$,其中$\sigma^2(x)$是$x$的平滑(但是未设定)函数,那么

(i) 在H_0^a下,$\bar{J}^a_{n,se}\overset{\text{def}}{=}n\bar{I}^a_{se}/\hat{S}_{n,a}\xrightarrow{d}N(0,1)$,其中

$$\hat{S}^2_{n,a}=2\sum_{i=1}^n\sum_{j=1}^n[p^K(X_i)'(P'P)^{-1}p^K(X_j)]^2\hat{u}_i^2\hat{u}_j^2$$

(ii) 在H_1^a下, 对任意$C_n=o(nK^{1/2})$,$P(\bar{J}^a_{n,se}>C_n)\to 1$。

定理 15.9 的证明在 15.6 节给出。

注意当误差是条件同方差时，我们可以用 $\sigma^2(x) = n^{-1}\sum_{i=1}^{n}\hat{u}_i^2$ 代替在 $S_{n,a}^2$ 内的 $\hat{u}_i^2$ 和 $\hat{u}_j^2$，并且 $S_{n,a}^2$ 简化为 $2\hat{\sigma}^4 K$。

Eubank 和 Hart(1992)提出了另一个关于正确参数回归函数形式的基于序列的检验。他们将参数原模型(如线性模型)嵌套进一个更一般的非参数序列回归模型，然后在模型的非参数部分进行序列系数的显著性检验。他们的方法的优点在于当原假定被拒绝时，估计的序列系数对于非线性项应该加入模型提供了指示。 474

15.4.2 关于可加部分线性模型的一致性检验

给定相对容易实施的非参数序列估计量服从的约束，比如服从可加性或者单调性约束，我们现在考虑构造基于序列方法的检验统计量。特别是，通过可加部分线性模型，我们关注检验模型的充分性这个问题。

我们在这一节讨论的检验统计量具有许多理想的性质，包括(i) 它避免了非参数地估计对立模型从而部分地避免了维数诅咒，(ii) 它可以检测到 $O_p(n^{-1/2})$ 阶的局部选择，并且(iii)它在计算上是简单的。

我们考虑以下可加部分线性原模型的一致性检验：

$$H_0^b: E(Y_i \mid X_i) = z_0(X_i)'\gamma + \sum_{l=1}^{L} m_l(X_{li}) \text{ a. s.}$$

$$\text{某些 } \gamma \in \mathcal{B}, \quad \sum_{l=1}^{L} m_l(\cdot) \in \mathcal{G} \tag{15.61}$$

其中$\mathcal{B}$是$\mathbb{R}^r$的一个紧子集且$\mathcal{G}$表示可加函数类。

备择假定 H_1^b 是 H_0^b 的否定，即

$$H_1^b: E(Y_i \mid X_i) \neq z_0(X_i)'\gamma + \sum_{l=1}^{L} m_l(X_{li}) \tag{15.62}$$

对任何 $\gamma \in \mathcal{B}$以及任何 $\sum_{l=1}^{L} m_l(\cdot) \in \mathcal{G}$，以正测度在一个集合上。

原假定 H_0^b 几乎确定等价于 $E(u_i \mid X_i) = 0$，其中 u_i 由上文定义。注意对所有 $M(\cdot) \in \mathcal{M}$，有界$\sigma(X_i)$测度函数类，当且仅当 $E[u_i \mathcal{M}(X_i)] = 0$ 时，几乎确定有 $E(u_i \mid X_i) = 0$。依照 Bierens 和 Ploberger(1997)，Stinchcombe 和 White (1998)以及 Stute(1997)，我们考虑具有以下形式的非条件矩检验，而不是式(15.61)的条件矩检验： 475

$$E[u_i \mathcal{H}(X_i, x)] = 0, \text{几乎所有 } x \in S \subset \mathbb{R}^q \tag{15.63}$$

其中$\mathcal{H}(\cdot,\cdot)$是一个加权函数的合适选择，该选择使得式(15.63)和式(15.61)等价。关于$\mathcal{H}$的具体条件见下面的假定 15.11(i)和(ii)。

我们假定加权函数$\mathcal{H}(\cdot,\cdot)$在$\mathcal{S}\times\mathcal{S}$上是有界的。Stinchcombe 和 White(1998)证明存在一类广泛的加权函数$\mathcal{H}(\cdot,\cdot)$使得式(15.63)几乎处处等价于$E(u_i|X_i)=0$。加权函数的选择包括指数函数$\mathcal{H}(X_i,x)=\exp(X_i'x)$，logistic 函数$\mathcal{H}(X_i,x)=1/[1+\exp(c-X_i'x)]$，其中$c\neq 0$，以及$\mathcal{H}(X_i,x)=\cos(X_i'x)+\sin(X_i'x)$；参见 Stinchcombe 和 White(1998)，Bierens 和 Ploberger(1997)对于此问题更进一步的讨论。

将$E[u_i\mathcal{H}(X_i,x)]$的样本对应值乘以$\sqrt{n}$，得到

$$J_n^0(x)=\sqrt{n}\left[\frac{1}{n}\sum_i u_i\mathcal{H}(X_i,x)\right]=\frac{1}{\sqrt{n}}\sum_i u_i\mathcal{H}(X_i,x) \qquad (15.64)$$

J_n^0可以看作在所有实分离空间$\mathcal{L}_2(\mathcal{S},\nu)$里取值的一个随机元素，在$\mathcal{S}$上的 Borel 可测函数$f$满足$\int_S f(x)^2\nu(\mathrm{d}x)<\infty$，它被赋予$L_2$范数$\|f\|_\nu^2=\int_S f(x)^2\nu(\mathrm{d}x)$。在20世纪60年代和70年代发展起来的 Hilbert 空间概率论，极大地简化了研究类似于$\|J_n^0\|_{L_2(\nu)}$的统计量的渐近分布问题，因为在$\mathcal{L}_2(\mathcal{S},\nu)$中取值的随机元素的中心极限定理的充分条件可以得到并且很容易检测。比如说，对于一个独立同分布$\mathcal{L}_2(\mathcal{S},\nu)$值元素的数列$\{Z_n(\cdot)\}_n$，Araujo 和 Giné(1980, p. 205)，van der Vaart 和 Wellner(1996, p. 50)，以及 Chen 和 White(1998)认为在L_2范数下($\mathcal{L}_2(\mathcal{S},\nu)$，$\|\cdot\|_{L_2(\nu)}$)，当且仅当$\int_S E[Z_1(x)^2]\nu(\mathrm{d}x)<\infty$时，$n^{-1/2}\sum_{i=1}^n Z_i(\cdot)$收敛于$\mathcal{Z}(\cdot)$。[①] 其中$\mathcal{Z}$是一个与$Z_1$具有相同协方差函数的高斯元素。我们将在下面的一个引理正式概括这个结论。

我们假定$\nu(\mathcal{S})<\infty$。因为我们将会只讨论$\mathcal{S}$是$\mathbb{R}^d$内的一个有界子集的情况，我们选择$\nu(\cdot)$作为在$\mathcal{S}$内的 Lebesgue 测度。那么$J_n^0(\cdot)$是一个在$\mathcal{L}_2(S,\nu)$内的 Hilbert 值随机元素。我们在下面的引理中提出了一个 Hilbert 值中心极限定理。

476

引理 15.1 令$Z_1(\cdot),\cdots,Z_n(\cdot)$为 Hilbert 值，独立同分布，是$\mathcal{L}_2(\mathcal{S},\nu)$上的零均值随机元素，满足$E[\|Z_i(\cdot)\|_\nu^2]<\infty$。然后，我们可以证明$n^{-1/2}\sum_{i=1}^n Z_i(\cdot)$弱收敛[①]于零均值高斯过程，其协方差(核)函数为

$$\Omega(x,x')=E[Z_i(x)Z_i(x)']$$

证明：见 Politis 和 Romano(1994)的定理 2.1，或者 van der Vaart 和 Wellner

① 如果对所有的实值有界连续函数h，有$E[h(\mathcal{Z}_n)]\to E[h(\mathcal{Z})]$，那么一个$H$值随机元素$\mathcal{Z}_n$序列弱收敛于$\mathcal{Z}$。

(1996,习题 1. 8. 5, p. 50)。注意 $E[\| Z_i(\cdot) \|_\nu^2] < \infty$ 是一个使得过程 $n^{-1/2}\sum_{i=1}^{n} Z_i(\cdot)$ 为紧的充分条件。□

使用引理 15.1 去检查 $J_n^0(\cdot)$是紧的是很直接的。令

$$Z_i(\cdot) = u_i \mathcal{H}(X_i, \cdot)$$

可得

$$\begin{aligned} E[\| Z_i(\cdot) \|_\nu^2] &= E\left\{\int u_i^2 [\mathcal{H}(X_i,x)]^2 \nu(\mathrm{d}x)\right\} \\ &= E\left\{\sigma^2(X_i)\int [\mathcal{H}(X_i,x)]^2 \nu(\mathrm{d}x)\right\} \\ &\leqslant E[\sigma^2(X_i)]\left\{C\int_{\mathcal{S}} \nu(\mathrm{d}x)\right\} < \infty \end{aligned}$$

其中 $\sigma^2(X_i) = E(U_i^2 | X_i)$。

于是,根据引理 15.1,我们知道 $J_n^0(\cdot)$在$\mathcal{L}_2(\mathcal{S},\nu,\|\cdot\|_\nu)$内弱收敛于 $J_\infty^0(\cdot)$。其中 $J_\infty^0(\cdot)$是一个高斯过程,以零为中心且协方差函数 Ω 为

$$\Omega(x,x') = E[Z_i(x)Z_i(x')] = E[\sigma^2(X_i)\mathcal{H}(X_i,x)\mathcal{H}(X_i,x')] \tag{15.65}$$

其中 x 和 $x' \in \mathcal{S}$。

因为 u_i 是不可观测的,我们必须用 $\hat{u}_i$ 代替 u_i,其定义由下面的式(15.67)给出,由此我们可以得出一个关于式(15.64)的可行版本为

$$\hat{J}_n(x) = \frac{1}{\sqrt{n}}\sum_i \hat{u}_i \mathcal{H}(X_i,x) \tag{15.66}$$

我们使用一个 K_l 函数的线性组合去逼近 $m_l(x_l)$ $(l=1,\cdots,L)$:

477

$$p_l^{K_l}(x_l) = (p_{l1}^{K_l}(x_l),\cdots,p_{lK_l}^{K_l}(x_l))'$$

也就是说,我们使用 $K = \sum_{l=1}^{L} K_l$ 函数的线性组合$(p_1^{K_1}(x_1)',\cdots,p_L^{K_L}(x_L)') \equiv p^K(x)'$去逼近一个可加函数 $\sum_{l=1}^{L} m_l(x_l)$。

我们使用 $\|\cdot\|$ 表示一般的欧几里得范数($\|\cdot\|_\nu$ 表示 L_2 范数)。我们假定:

假定 15.11

(i) 加权函数$\mathcal{H}(X_i,x) = w(X_i'x)$,$w(\cdot)$是一个解析的非多项式函数。①

(ii) $\mathcal{H}(\cdot,\cdot)$在$\mathcal{S}\times\mathcal{S}$上是有界的并且满足 Lipschitz 条件,即对于所有的

① 解析函数是指在它域内的每一个点局部等于它的 Talor 展开式,比如 exp(·)、logistic、双曲切线、正弦和余弦函数等。

$x_1, x_2 \in \mathcal{S}$, $|\mathcal{H}(X_i, x_1) - \mathcal{H}(X_i, x_2)| \leqslant G(X_i) \| x_1 - x_2 \|$，有 $E[G^2(X_i)] < \infty$。

(iii) $v(\cdot)$是 Lebesgue 测度。

令 p_l 和 P 为式(15.20)所定义的且令 Z_0 为 $n \times r$ 矩阵，其第 i 行由 $z_0(X_i)'$ 给出。然后用向量—矩阵表示法把式(15.61)写为

$$\mathcal{Y} = Z_0\gamma + m + U = Z_0\gamma + P\beta + (m - P\beta) + U = \mathcal{X}\alpha + (m - P\beta) + U$$

其中$\mathcal{X} = (Z_0, P)$, $\alpha = (\gamma', \beta')'$; Y 和 U 都是 $n \times 1$ 向量，其第 i 个分量由 Y_i 和 u_i 分别给出；并且 m 是 $n \times 1$ 向量，其第 i 个分量为 $m_i = \sum_{l=1}^{L} m_l(X_{li})$。$P$ 是 $n \times K$ 维的，且 $\beta = \beta_m$是 $K \times 1$ 向量，满足假定 15.2(iii)($f = m$)。

我们用最小二乘法通过$\mathcal{Y}$对$\mathcal{X}$回归来估计 $\alpha = (\gamma', \beta')'$，得到

$$\hat{\alpha} = \begin{bmatrix} \hat{\gamma} \\ \hat{\beta} \end{bmatrix} = (\mathcal{X}'\mathcal{X})^{-}\mathcal{X}'\mathcal{Y}$$

478 其中$(\mathcal{X}'\mathcal{X})^{-}$是$(\mathcal{X}'\mathcal{X})$的广义逆。Li(2000)证明 $\hat{\gamma} - \gamma = O_p(n^{-1/2})$。同样，由 Andrews 和 Whang(1990)以及 Newey(1995,1997)的结果可知

$$\hat{m}(x) - m(x) = O_p\left((K/n)^{1/2} + \sum_{l=1}^{L} K_l^{-\delta_l}\right)$$

其中 $\hat{m}(x) = p^K(x)'\hat{\beta}$，于是，我们用下式估计 u_i：

$$\hat{u}_i = Y_i - z_0(X_i)'\hat{\gamma} - p^K(X_i)'\hat{\beta} \tag{15.67}$$

我们对于 H_0^b 的检验统计量是基于

$$\hat{J}_n(x) = \frac{1}{\sqrt{n}} \sum_i \mathcal{H}(X_i, x)\hat{u}_i$$

其中 $\hat{u}_i$ 由式(15.67)给出。由 $\hat{J}_n(x)$我们可以构建一个 Cramer-von Mises(CM)类型的统计量来检验 H_0^b：

$$\mathrm{CM}_n = \int [\hat{J}_n(x)]^2 F_n(\mathrm{d}x) = \frac{1}{n} \sum_i [\hat{J}_n(X_i)]^2$$

其中 $F_n(\cdot)$是 $X_1, \cdots, X_n$ 的经验分布函数。

接下来的定理建立了在 H_0^b 假定下的 $\hat{J}_n(x)$和 CM_n 的弱收敛性。

定理 15.10 如果假定 15.6、15.7 和 15.11 成立。那么，在 H_0^b 下，

(i) $\hat{J}_n(\cdot)$在$\mathcal{L}_2(\mathcal{S}, \nu, \|\cdot\|_\nu)$上弱收敛于 $J_\infty(\cdot)$，J_∞是一个高斯过程，具有零均值且协方差函数由 $\Sigma_1(x, x') = E[\sigma^2(X_i)\eta_i(x)\eta_i(x')]$给出，其中 $\eta_i(x) = \mathcal{H}(X_i, x) - \phi_i(x) - \psi_i(x)$，且

$$\phi_i(x) = E_{\mathcal{G}}[\mathcal{H}(X_i, x)]$$

$$\psi_i(x) = E[\mathcal{H}(X_i,x)\varepsilon_i]\{E[\varepsilon_i\varepsilon_i']\}^{-1}\varepsilon_i$$
$$\varepsilon_i = z_0(X_i) - E_{\mathcal{G}}(z_0(X_i))$$

(ii) CM_n 依分布收敛于 $\int[J_\infty(x)]^2F(\mathrm{d}x)$，其中 $F(\cdot)$ 是 X_i 的累积分布函数。

定理15.10(i)的证明在15.6节给出。

正如 Bierens 和 Ploberger(1997)以及 Chen 和 Fan(1999)所倡导的，人们可以证明 $\int[J_\infty(x)]^2F(\mathrm{d}x)$ 可以写成加权的(独立)χ_1^2 随机变量的无限总和，其权数由 (X_i,Y_i) 的未知分布决定。因此，得到临界值是困难的。我们建议对于 CM_n 的零极限分布使用一个基于残差的原始自举法去逼近临界值，即 $u_i^* = [(1-\sqrt{5})/2]\hat{u}_i$，概率为 $(1+\sqrt{5})/[2\sqrt{5}]$；$u_i^* = [(1+\sqrt{5})/2]\hat{u}_i$，概率为 $(\sqrt{5}-1)/[2\sqrt{5}]$。注意 u_i^* 满足 479

$$E^*(u_i^*) = 0,\quad E^*(u_i^{*2}) = \hat{u}_i^2,\quad E^*(u_i^{*3}) = \hat{u}_i^3$$

其中 $E^*(\cdot)=E(\cdot\mid\mathcal{W}_n)$ 且 $\mathcal{W}_n=\{Y_i,X_i\}_{i=1}^n$。从 $\{u_i^*\}_{i=1}^n$，我们可以通过以下原模型生成 Y_i^*：

$$Y_i^* = z_0(X_i)'\hat{\gamma} + p^K(X_i)'\hat{\beta} + u_i^*,\quad i=1,\cdots,n$$

然后，使用自举样本 $\{(Y_i^*,X_i)\}_{i=1}^n$，可得

$$\begin{bmatrix}\hat{\gamma}^*\\ \hat{\beta}^*\end{bmatrix} = (\mathcal{X}'\mathcal{X})^-\mathcal{X}'\mathcal{Y}^*$$

其中 $\mathcal{X}=(Z_0,P)$ 和 $\mathcal{Y}^*$ 是 $n\times 1$ 向量，其第 j 个元素是 Y_i^*。自举残差由 $\hat{u}_i^* = Y_i^* - z_0(X_i)'\hat{\gamma}^* - p^K(X_i)'\hat{\beta}^*$ 给出，且自举统计量 $\hat{J}_n^*(x)$ 是通过把 $\hat{J}_n(x)$ 中的 $\hat{u}_i$ 以 $\hat{u}_i^*$ 替代得到，即

$$\hat{J}_n^*(x) = \frac{1}{\sqrt{n}}\sum_i \hat{u}_i^*\,\mathcal{H}(X_i,x)$$

使用 $\hat{J}_n^*(x)$，我们可以计算 CM_n 统计量的自举形式，即

$$CM_n^* = \frac{1}{n}\sum_i[\hat{J}_n^*(X_i)]^2$$

Li, Hsiao 和 Zinn(2003)证明了 CM_n^* 与 CM_n 的零极限分布同样收敛于相同的极限分布。Li, Hsiao 和 Zinn 的模拟结果表明自举统计量 CM_n^* 在有限样本应用当中表现良好。

15.5 其他基于序列的检验

Li, Hsiao 和 Zinn(2003)也考虑了用基于序列的方法去检验一个不具有参数部分线性成分的可加回归模型的原假定,这可以看成是一个特殊情形,即检验一个 $\gamma=0$ 的可加部分线性模型。Li,Hsiao 和 Zinn 进一步考虑了检验一个部分线性模型的原假定,这也可以看成是一个特例,即检验一个可加局部线性模型,但有一个($L=1$)非参数函数(而非适用于 $L>1$ 个非参数可加函数)。理论上,可以对所有类型的假定构造基于序列的检验,包括非参数遗漏变量检验和对两个未知密度的相等性检验。

480

Donald,Imbens 和 Newey(2003)提出了一个新的方法去构建一个基于一个经验似然函数的序列估计的一致条件矩检验。他们的检验统计量基于出现在条件矩约束中的有限维参数的有效估计。其方法允许内生回归元并且也允许大量的矩条件。

15.6 证　明

依据 Newey(1997)给出的论证,我们假定在假定 15.2 中 $B=I$。因此 $p^K(\cdot)=P^K(\cdot)$,因为所有的非参数序列估计量对于 $p^K(\cdot)$ 的非奇异变换来说都是不变的。同样我们假定 $Q\overset{\text{def}}{=}E[p^K(X_i)p^K(X_i)']=I$。这是因为对于 Q^{-1} 的对称平方根 $Q^{-1/2}$,$Q^{-1/2}p^K(\cdot)$ 是 $p^K(\cdot)$ 的一个非奇异变换,并且,由假定 15.2(ii)可以很容易证明 $\tilde{\zeta}_0(K)\overset{\text{def}}{=}\sup_{x\in\mathcal{S}}\|Q^{-1/2}p^K(\cdot)\|\leqslant C\zeta_0(K)$。进一步,如果我们把 $p^K(\cdot)$ 换成 $\bar{p}^K(\cdot)\equiv Q^{-1/2}p^K(\cdot)$ 并且定义 $\bar{\beta}=Q^{1/2}\beta$,那么,假定 15.2(iii)满足,因为 $|g(\cdot)-p^K(\cdot)'\beta|=|f(\cdot)-\bar{p}^K(\cdot)'\bar{\beta}|$。于是,当 $p^K(\cdot)$ 换成 $Q^{-1/2}p^K(\cdot)$ 时,所有的假定成立。

15.6.1 定理 15.1 的证明

我们在这里只证明定理 15.1(i)。定理 15.1(ii)和(iii)可以用同样的方法证明。

481

(i)的证明:利用三角不等式($\|A+B\|\leqslant\|A\|+\|B\|$),令 $\mathbf{1}_n$ 表示大于 1/2 的 $\hat{Q}$ 的最小特征值的示性函数,由引理 15.3(见下文)和假定 15.2,

$$\mathbf{1}_n\int[\hat{g}(x)-g(x)]^2\mathrm{d}F(x)$$

$$= \mathbf{1}_n\int [p^K(x)'(\hat{\beta}-\beta)+p^K(x)'\beta-g(x)]^2\mathrm{d}F(x)$$

$$\leqslant 2\mathbf{1}_n\int[p^K(x)'(\hat{\beta}-\beta)]^2\mathrm{d}F(x)+2\mathbf{1}_n\int[p^K(x)'\beta-g(x)]^2\mathrm{d}F(x)$$

$$= 2\mathbf{1}_n\|\hat{\beta}-\beta\|^2+2\mathbf{1}_n\int[p^K(x)'\beta-g(x)]^2\mathrm{d}F(x)$$

$$\leqslant O_p(K/n+K^{-2\alpha})+O_p(K^{-2\alpha})=O_p(K/n+K^{-2\alpha})$$

在上面的第二个等式中，我们使用

$$\int[p^K(x)'(\hat{\beta}-\beta)]^2\mathrm{d}F(x)=(\hat{\beta}-\beta)'\int p^K(x)p^K(x)'\mathrm{d}F(x)(\hat{\beta}-\beta)$$

$$=\|\hat{\beta}-\beta\|^2$$

因为$\int p^K(x)p^K(x)'\mathrm{d}F(x)=E[p^K(X)p^K(X)]=I_K$。 □

现在我们证明在定理 15.1 的证明中使用到的引理。

引理 15.2　$E[\|\hat{Q}-I\|^2]=O(\zeta_0(K)^2K/n)(Q=I_K)$。

证明：回忆一下，对于任意的矩阵 A，$\|A\|^2$ 是由 $\|A\|^2=\mathrm{tr}(A'A)$ 定义的。对于一个 $K\times K$ 矩阵，令 A_{jl}表示出现在 A 中第 j 行和第 l 列的元素，我们有

$$\mathrm{tr}(A'A)=\sum_{j=1}^{K}\sum_{l=1}^{K}A_{jl}^2$$

令 δ_{jl}表示$(I_K)_{jl}$，注意假定 $Q=I$ 意味着　482

$$E[p_{jK}(X_i)p_{lK}(X_i)]=\delta_{jl}$$

可得

$$E[\|\hat{Q}-I\|^2]=E[\|(P'P/n)-I\|^2]$$

$$=\sum_{j=1}^{K}\sum_{l=1}^{K}E[\{n^{-1}\sum_{i=1}^{n}p_{jK}(X_i)p_{lK}(X_i)-\delta_{js}\}^2]$$

$$\left(\text{由 }\|A\|^2=\mathrm{tr}(A'A)=\sum_{j=1}^{K}\sum_{l=1}^{K}A_{jl}^2\right)$$

$$=\sum_{j=1}^{K}\sum_{l=1}^{K}n^{-1}E\{[p_{jK}(X_i)p_{lK}(X_i)-\delta_{jl}]^2\}$$

$$\left(\text{因为 }E[p_{jK}(X_i)p_{lK}(X_i)]=\delta_{jl}\right)$$

$$\leqslant\sum_{j=1}^{K}\sum_{l=1}^{K}E[p_{jK}(X_i)^2p_{lK}(X_i)^2]/n$$

$$\left(\text{当 }A=p_{jK}(X_i)p_{lK}(X_i)\text{ 时，}\mathrm{var}(A)\leqslant E(A^2)\right)$$

$$=E[\|p^K(X_i)\|^2\sum_{j=1}^{K}p_{jK}(X_i)^2]/n$$

$$\leqslant \zeta_0(K)^2 E[\sum_{j=1}^{K} p_{jK}(X_i)^2]/n$$

(由假定 15.2)

$$= \zeta_0(K)^2 E\{\mathrm{tr}[p^K(X_i)p^K(X_i)']\}/n$$

$$= \zeta_0(K)^2 \mathrm{tr}(I_K)/n = \zeta_0(K)^2 K/n$$

注意引理 15.2 意味着 $\|\hat{Q}-I\|^2 = O_p(\zeta_0(K)\sqrt{K/n}) = o_p(1)$。同样,因为 $\hat{Q}-I$ 的最小特征值是受 $\|\hat{Q}-I\|$ 约束的,这意味着 $\hat{Q}$ 的最小特征值依概率收敛于 1,即 $P(\mathbf{1}_n=1)\to 1$。 □

引理 15.3 $\mathbf{1}_n\|\hat{\beta}-\beta\| = O_p(K^{1/2}/n^{1/2}+K^{-\alpha})$。

483

证明:令 $u=(u_1,\cdots,u_n)'$,可得

$$\begin{aligned}E[\mathbf{1}_n\|\hat{Q}^{-1/2}P'u/n\|^2 \mid X] &= \mathbf{1}_n E[u'P\hat{Q}^{-1}P'u \mid X]/n^2 \\ &= \mathbf{1}_n E[u'P(P'P)^{-1}P'u \mid X]/n \\ &= \mathbf{1}_n E[\mathrm{tr}\{u'P(P'P)^{-1}P'u\} \mid X]/n \\ &= \mathbf{1}_n \mathrm{tr}\{P(P'P)^{-1}P'E[uu' \mid X]\}/n \\ &\leqslant C\mathbf{1}_n \mathrm{tr}\{P(P'P)^{-1}P'\}/n \\ &= \mathbf{1}_n CK/n\end{aligned}$$

因此,由 Markov 不等式,

$$\begin{aligned}\mathbf{1}_n\|\hat{Q}^{-1}P'u/n\| &= \mathbf{1}_n\{(u'P/n)\hat{Q}^{-1/2}\hat{Q}^{-1}\hat{Q}^{-1/2}(P'u/n)\}^{1/2} \\ &\leqslant O_p(1)\mathbf{1}_n\|\hat{Q}^{-1/2}P'u/n\| \\ &= O_p((K/n)^{1/2})\end{aligned} \tag{15.68}$$

令 $\beta=\beta_K$ 为假定 15.2 中给出的,并且令 G 为一个 $n\times 1$ 矩阵,其第 i 行为 $g(X_i)$。那么,注意 $\mathbf{1}_n P(P'P)^{-1}P'$ 是等幂的,则

$$\begin{aligned}&\mathbf{1}_n\|\hat{Q}^{-1}P'(G-P\beta)/n\| \\ &= \mathbf{1}_n[(G-P\beta)'P\hat{Q}^{-1/2}\hat{Q}^{-1}\hat{Q}^{-1/2}P'(G-P\beta)/n^2]^{1/2} \\ &\leqslant O_p(1)\mathbf{1}_n[(G-P\beta)'P\hat{Q}^{-1/2}\hat{Q}^{-1/2}P'(G-P\beta)/n^2]^{1/2} \\ &= O_p(1)\mathbf{1}_n[(G-P\beta)'P(P'P)^{-1}P'(G-P\beta)/n]^{1/2} \\ &\leqslant O_p(1)\mathbf{1}_n[(G-P\beta)'(G-P\beta)/n]^{1/2} \\ &= O_p(K^{-\alpha})\end{aligned} \tag{15.69}$$

其中,在上面的最后一个不等式中,我们使用了如果 A 是一个等幂矩阵,则 $z'Az\leqslant z'z$,且最后一个等式来自假定 15.2。

因此，通过使用三角不等式，可得

$$\begin{aligned}\|\mathbf{1}_n(\hat{\beta}-\beta)\| &= \|\mathbf{1}_n\hat{Q}^{-1}P'y/n-\mathbf{1}_n\hat{Q}^{-1}P'P\beta/n\| \\ &= \|\mathbf{1}_n\hat{Q}^{-1}P'(y-G)/n+\mathbf{1}_n\hat{Q}^{-1}P'(G-P\beta)/n\| \\ &\leqslant \mathbf{1}_n\|\hat{Q}^{-1}P'u/n\| + \mathbf{1}_n\|\hat{Q}^{-1}P'(G-P\beta)/n\| \\ &= O_p((K/n)^{1/2}+K^{-\alpha})\end{aligned}$$

其中最后的等式根据式(15.68)和式(15.69)推出。 □

15.6.2　定理 15.3 的证明

484

回忆 $\theta(X_i)=E(Z_i|X_i)$，$v_i=Z_i-\theta(X_i)$，我们将使用下面的缩写记号：$\theta_i=\theta(X_i)$，$g_i=g(X_i)$ 和 $v_i=Z_i-\theta_i$。

为了避免在向量—矩阵变量中引入过多的记号，我们将使用没有下标的同样的记号来表示一个向量或者一个矩阵。例如，θ 是 $n\times r$ 矩阵，其第 i 行为 $\theta(Z_i)'$。这种惯例适用于 g、v、u 等。

回忆我们定义 $\tilde{A}$ 为 $\tilde{A}=P(P'P)^{-}P'A$。这个定义适用于任意一个下面考虑到的 $n\times 1$ 或者 $n\times r$ 矩阵。例如，$\tilde{\theta}=P(P'P)^{-}P'\theta$，而且 $\tilde{u}$ 和 $\tilde{v}$ 被同样地定义。

由 $X_i=\theta_i+v_i$，我们得到 $\tilde{Z}_i=\tilde{\theta}_i+\tilde{v}_i$。或者说，在向量—矩阵表示法中，$Z=\theta+v$，$\tilde{Z}=\tilde{\theta}+\tilde{v}$。于是，我们可得

$$Z-\tilde{Z}=(\theta-\tilde{\theta})+v-\tilde{v} \tag{15.70}$$

对于标量或者列向量 A_i 和 B_i，定义 $S_{A,B}=n^{-1}\sum_i A_iB_i'$。同样，我们令 $S_{A,A}=S_A$。

注意如果 $S_{Z-\tilde{Z}}^{-1}$ 存在，那么我们得到

$$\sqrt{n}(\hat{\gamma}-\gamma)=S_{Z-\tilde{Z}}^{-1}\sqrt{n}S_{Z-\tilde{Z},g-\tilde{g}+u-\tilde{u}}$$

我们将通过证明下列式子来证明定理 15.3。

(i) $S_{Z-\tilde{Z}}=\Phi+o_p(1)$（因此 $S_{Z-\tilde{Z}}$ 是渐近非奇异的）；

(ii) $S_{Z-\tilde{Z},g-\tilde{g}}=o_p(n^{-1/2})$；

(iii) $S_{Z-\tilde{Z},\tilde{u}}=o_p(n^{-1/2})$；

(iv) $\sqrt{n}S_{Z-\tilde{Z},\tilde{u}}\to N(0,\Omega)$ 依分布收敛。

这些在下文将被证明。

(i)的证明：$S_{Z-\tilde{Z}}=\Phi+o_p(1)$。由式(15.7)，可得

$$S_{Z-\tilde{Z}}=S_{v+(\theta-\tilde{\theta})-\tilde{v}}=S_v+S_{(\theta-\tilde{\theta})-\tilde{v}}+2S_{v,(\theta-\tilde{\theta})-\tilde{v}}$$

首先，根据大数定律的性质，$S_v = n^{-1}\sum_i v_i v_i' = \Phi + o_p(1)$。

其次，由引理 15.7 以及引理 15.8(i)可得

$$S_{(\theta-\tilde{\theta})-\tilde{v}} \leqslant 2\{S_{(\theta-\tilde{\theta})} + S_{\tilde{v}}\} = o_p(1)$$

最后，由以上结果可知

$$S_{v,(\theta-\tilde{\theta})-\tilde{v}} \leqslant \{S_v S_{(\theta-\tilde{\theta})-\tilde{v}}\}^{1/2} = \{O_p(1)o_p(1)\}^{1/2} = o_p(1) \qquad \square$$

485 (ii)的证明 $S_{Z-\tilde{Z},g-\tilde{g}} = O_p(K^{-\delta}) = o_p(n^{-1/2})$。利用式(15.70)，

$$S_{Z-\tilde{Z},g-\tilde{g}} = S_{v+(\theta-\tilde{\theta})-\tilde{v},g-\tilde{g}} = S_{v,g-\tilde{g}} + S_{\theta-\tilde{\theta},g-\tilde{g}} - S_{\tilde{v},g-\tilde{g}}$$

并且我们分别考虑等式右边的三项。

(i) $E[\| S_{v,g-\tilde{g}} \|^2 | X] = n^{-2}\mathrm{tr}[(g-\tilde{g})(g-\tilde{g})'E(vv'|X)]$

$$\begin{aligned} &\leqslant Cn^{-2}\mathrm{tr}[(g-\tilde{g})(g-\tilde{g})'] \\ &= Cn^{-1}\mathrm{tr}(S_{g-\tilde{g}}) \\ &= n^{-1}O_p(K^{-2\delta}) = o_p(n^{-1}) \end{aligned}$$

因此，$S_{v,g-\tilde{g}} = o_p(n^{-1/2})$。

(ii) 由引理 15.7，$S_{\theta-\tilde{\theta},g-\tilde{g}} \leqslant \{S_{\theta-\tilde{\theta}} S_{g-\tilde{g}}\}^{1/2} = O_p(K^{-2\delta})$。

(iii) 由引理 15.7 和引理 15.8(i)，$S_{\tilde{v},g-\tilde{g}} \leqslant \{S_{\tilde{v}} S_{g-\tilde{g}}\}^{1/2} = o_p(1)O_p(K^{-\delta})$。

$\square$

(iii)的证明：$S_{Z-\tilde{Z},\tilde{u}} = o_p(n^{-1/2})$。利用式(15.70)，

$$S_{Z-\tilde{Z},\tilde{u}} = S_{v+(\theta-\tilde{\theta})-\tilde{v},\tilde{u}} = S_{v,\tilde{u}} + S_{\theta-\tilde{\theta},\tilde{u}} - S_{\tilde{v},\tilde{u}}$$

我们分别考虑这三项。

(i) 由引理 15.8(i)，

$$\begin{aligned} E[\| S_{v,\tilde{u}} \|^2 \mid X,Z] &= n^{-2}\mathrm{tr}[P(P'P)^- P'vv'P(P'P)^- P'E(uu' \mid X,Z)] \\ &\leqslant Cn^{-2}\mathrm{tr}[P(P'P)^- P'vv'P(P'P)^- P'] \\ &= Cn^{-2}\mathrm{tr}[\tilde{v}\tilde{v}'] = Cn^{-1}\mathrm{tr}(S_{\tilde{v}}) = O(K/n^2) \end{aligned}$$

因此，$S_{v,\tilde{u}} = O_p(\sqrt{K}/n)$。

(ii) 由引理 15.7 和引理 15.8(ii)，$S_{\theta-\tilde{\theta},\tilde{u}} \leqslant \{S_{\theta-\tilde{\theta}} S_{\tilde{u}}\}^{1/2} = O_p(K^{-\delta})$
$O_p(\sqrt{K}/\sqrt{n})$。

(iii) 由引理 15.8(i)和(ii)，$S_{\tilde{v},\tilde{u}} \leqslant \{S_{\tilde{v}} S_{\tilde{u}}\}^{1/2} = O_p(K/n)$。 $\square$

(iv)的证明：$\sqrt{n}S_{Z-\tilde{Z},\tilde{u}} \to N(0,\Omega)$依分布收敛。

$$S_{Z-\tilde{Z},u} = S_{v+(\theta-\tilde{\theta})-\tilde{v},\tilde{u}} = S_{v,u} + S_{\theta-\tilde{\theta},u} - S_{\tilde{v},u}$$

我们分别考虑这三项。

486 (i) 由 Lindeberg-Levy 中心极限定理，$\sqrt{n}S_{v,u} = n^{-1/2}\sum_i v_i u_i \to N(0,\Omega)$ 依分

布收敛。

(ii) 由引理 15.7,

$$E[\| S_{\theta-\tilde{\theta},u} \|^2 \mid X] = n^{-2}\mathrm{tr}[(\theta-\tilde{\theta})(\theta-\tilde{\theta})'E(uu' \mid X)]$$
$$\leqslant Cn^{-2}\mathrm{tr}[(\theta-\tilde{\theta})(\theta-\tilde{\theta})'/n]$$
$$= Cn^{-1}\mathrm{tr}(S_{\theta-\tilde{\theta}}) = o_p(n^{-1})$$

因此,$S_{\theta-\tilde{\theta},u} = O_p(n^{-1/2})$。

用与上面的(ii)完全一样的论证,可得

(iii) 由引理 15.8(i),$E[\| S_{\tilde{v},u} \|^2 \mid X] \leqslant Cn^{-1}\mathrm{tr}(S_{\tilde{v}}) = o_p(n^{-1})$。因此,$S_{\tilde{v},u} = o_p(n^{-1/2})$,且以上(i)—(iv)意味着

$$\sqrt{n}(\hat{\gamma}-\gamma) = \Phi^{-1}N(0,\Omega) + o_p(1) \to N(0,\Phi^{-1}\Omega\Phi^{-1})\text{依分布收敛}$$

$\hat{\Sigma} = \Sigma + o_p(1)$ 的证明:$\hat{\Sigma} = \hat{\Phi}^{-1}\hat{\Omega}\hat{\Phi}^{-1}$, $\hat{\Phi} \equiv S_{X-\tilde{X}} = \Phi + o_p(1)$ 上面已证明。下面我们提供一个证明 $\hat{\Omega} = \Omega + o_p(1)$ 的概要,因为详细的证明与 $\hat{\Phi} = \Phi + o_p(1)$ 的证明是非常相似的。

使用 $\hat{\gamma}-\gamma = O_p(n^{-1/2})$ 和 $\hat{g}(X_i) - g(X_i) = o_p(1)$,易知 $\hat{u}_i = u_i + o_p(1)$。同样,由引理 15.7 以及引理 15.8(ii),可知 $\theta_i - \tilde{\theta}_i = o_p(1)$,$\tilde{v}_i = o_p(1)$,$\tilde{\eta}_i = o_p(1)$。于是,由式(15.70)可知,$Z_i - \tilde{Z}_i = v_i + (\theta_i - \tilde{\theta}_i) - \tilde{v}_i = v_i + o_p(1)$。这些结果推出 $\hat{\Omega} = n^{-1}\sum_i \hat{u}_i^2(Z_i - \tilde{Z}_i)(Z_i - \tilde{Z}_i)' = n^{-1}\sum_{i=1}^{n} u_i^2 v_i v_i' + o_p(1) = \Omega + o_p(1)$(根据大数定律的性质)。 □

我们使用 $\mathbf{1}_n$ 表示一个示性函数,当 $(P'P)$ 可逆时这个函数取值为 1,反之为 0。我们将只在引理 15.5 的证明中明确地使用示性函数 $\mathbf{1}_n$,而在引理 15.6 至引理 15.8(iii)的证明中省略它以简化记号。不论我们何时使用 $(P'P)^{-1}$,它应该被理解为 $\mathbf{1}_n(P'P)^{-1}$,并且因为 $P(\mathbf{1}_n = 1) \to 1$ 依概率收敛,所以我们将经常省略示性函数 $\mathbf{1}_n$。

引理 15.4　$\hat{Q} - I = O_p(\zeta_0(K)\sqrt{K}/\sqrt{n})$,其中 $\hat{Q} = (P'P/n)$。

证明:参见 Newey(1997, pp. 161—162)关于定理 1 的证明。 □

引理 15.5　$\| \tilde{\beta}_f - \beta_f \| = O_p(K^{-\alpha})$,其中 $\tilde{\beta}_f = (P'P)^{-}P'f$,$\beta$ 满足假定 15.2,$f = g$ 或 $f = \theta$。

证明:由引理 15.2、假定 15.2 以及 $P(P'P)^{-1}P'$ 是等幂的事实,

487

$$\mathbf{1}_n \| \tilde{\beta}_f - \beta_f \| = \mathbf{1}_n \| (P'P)^{-}P'(f - P\beta_f) \|$$
$$= \mathbf{1}_n \{(f - P\beta_f)'P(P'P)^{-1}(P'P/n)^{-1}P'(f - P\beta_f)/n\}^{1/2}$$

$$= \mathbf{1}_n O_p(1) \{(f - P\beta_f)'P(P'P)^- P'(f - P\beta_f)/n\}^{1/2}$$
$$\leqslant O_p(1) \{(f - P\beta_f)'(f - P\beta_f)/n\}^{1/2} = O_p(K^{-\alpha})$$

最后，$\|\tilde{\beta}_f - \beta_f\| = O_p(K^{-\alpha})$，因为 $P(\mathbf{1}_n = 1) \to 1$ 。 □

引理 15.6

$$(P'\eta/n) = O_p(\zeta_0(K)/\sqrt{n}) = o_p(1)$$

其中 $\eta = u$ 或 $\eta = v$。

证明：注意 $E[P_i\eta_i] = 0$，可得

$$E \|P'\eta/n\|^2 = n^{-2}\{\sum_i \sum_j E[P_i'P_j\eta_i\eta_j]\}$$
$$= n^{-2}\{\sum_i E[P_iP_i'\eta_i^2] + \sum_i \sum_{j\neq i} E[P_i'\eta_i]E[P_j\eta_j]\}$$
$$= n^{-2}\sum_i E[P_i'P_i\eta_i^2]$$
$$\leqslant Cn^{-1}E[P_i'P_i] = O_p((\zeta_0(K))^2/n)$$

因此，$(P'\eta/n) = O_p(\zeta_0(K)/\sqrt{n})$。

引理 15.7 $S_{f-\tilde{f}} = O_p(K^{-2\alpha}) = o_p(n^{-1/2})$，其中 $f = g$ 或 $f = \theta$。

证明：注意 $\tilde{f} \equiv P\tilde{\beta}_f$，由假定 15.2 和引理 15.2 以及引理 15.5，可得

$$S_{f-\tilde{f}} = n^{-1} \|f - \tilde{f}\|^2 \leqslant 2n^{-1}\{\|f - P\beta\|^2 + \|P(\beta_f - \tilde{\beta}_f)\|^2\}$$
$$= O(K^{-2\alpha}) + (\beta_f - \tilde{\beta}_f)'(P'P/n)(\beta_f - \tilde{\beta}_f)$$
$$= O(K^{-2\alpha}) + O_p(1) \|\beta_f - \tilde{\beta}_f\|^2 = O_p(K^{-2\alpha})$$

引理 15.8

(i) $S_{\tilde{v}} = O_p(K/n)$

(ii) $S_{\tilde{u}} = O_p(K/n)$

488

(i)的证明：与引理 15.2 的证明一样，可知

$$E[S_{\tilde{v}} \mid Z] = n^{-1}E\{v'P(P'P)^{-1}P'v \mid Z\}$$
$$= n^{-1}\mathrm{tr}[P(P'P)^{-1}P'E(v'v \mid Z)]$$
$$\leqslant Cn^{-1}\mathrm{tr}[P(P'P)^- P'] = O_p(K/n)$$

这意味着 $S_{\tilde{v}} = O_p(K/n)$。

(ii)的证明：与引理 15.8(i)的证明相似。 □

15.6.3 定理 15.6 的证明

回忆 $\theta(X_i, Z_i) = E[W_i \mid X_i, Z_i]$，$m(X_i, Z_i) = E_{\mathcal{G}}(W_i) = E_{\mathcal{G}}(\theta(X_i, Z_i))$ ，并

且 $\varepsilon_i = W_i - m(Z_i, X_i)$。定义 $v_i = W_i - \theta_i(X_i, Z_i)$ 和 $\eta_i = \theta(Z_i, X_i) - m(X_i, Z_i)$。我们使用下面的缩写记号：$\theta_i = \theta(X_i, Z_i)$，$g_i = Z_i'\beta(X_i)$，$m_i = m(X_i, Z_i)$。于是，$v_i = W_i - \theta_i$，$\varepsilon_i = \theta_i + v_i - m_i$，$\eta_i = \theta_i - m_i$。最后，没有下标的变量表示矩阵，例如，$\theta = (\theta_1, \cdots, \theta_n)'$是 $n \times 1$ 维的。

同样回忆对于任意的有 n 行的矩阵 A，我们定义 $\tilde{A} = P(P'P)^{-}P'A$（$P$ 是由下面的式(15.47)定义的）。把这个定义应用到 θ、m、g、η、u、v，我们得到 $\tilde{\theta}$、$\tilde{m}$、$\tilde{g}$、$\tilde{\eta}$、$\tilde{u}$、$\tilde{v}$。

因为 $W_i = \theta_i + v_i$，$\theta_i = m_i + \eta_i$，我们得到 $W_i = \eta_i + v_i + m_i$ 和 $\tilde{W}_i = \tilde{\eta}_i + \tilde{v}_i + \tilde{m}_i$。用矩阵表示，$\mathcal{W} = \eta + v + m$，$\tilde{\mathcal{W}} = \tilde{\eta} + \tilde{v} + \tilde{m}$。因此，可得

$$\mathcal{W} - \tilde{\mathcal{W}} = \eta + v + (m - \tilde{m}) - \tilde{v} - \tilde{\eta} \tag{15.71}$$

对于标量和列向量 A_i 和 B_i，我们定义 $S_{A,B} = n^{-1}\sum_i A_i B_i'$ 和 $S_A = S_{A,A}$。我们还定义一个标量函数 $\bar{S}_A = n^{-1}\sum_i A_i A_i'$，是 S_A 的对角线元素的和。利用 $ab \leqslant (a^2 + b^2)/2$，易知 $S_{A,B}$ 的每一个元素是小于或者等于 $\bar{S}_A + \bar{S}_B$ 的。当我们评估 $S_{A,B}$ 的概率阶数时，我们通常记作 $S_{A,B} \leqslant \bar{S}_A + \bar{S}_B$。标量边界 $\bar{S}_A + \bar{S}_B$ 约束 $S_{A,B}$ 内的每一个元素。因此，如果 $\bar{S}_A + \bar{S}_B = O_p(a_n)$（对某些正的序列 a_n），那么 $S_{A,B}$ 的每个元素至多是 $O_p(a_n)$，这意味着 $S_{A,B} = O_p(a_n)$。同样，使用柯西-施瓦茨不等式，可得 $S_{A,B} \leqslant (\bar{S}_A \bar{S}_B)^{1/2}$。再一次，标量约束了 $S_{A,B}$ 内的所有元素。

注意，如果 $S^{-1}_{W-\tilde{W}}$ 存在，那么由式(15.51)和式(15.52)，可得

489

$$\begin{aligned}
\sqrt{n}(\hat{\gamma} - \gamma) &= \left[n^{-1}\sum_i (W_i - \tilde{W}_i)(W_i - \tilde{W}_i)'\right]^{-1} \\
&\quad \times \sqrt{n}\left\{n^{-1}\sum_i (W_i - \tilde{W}_i)(g_I - \tilde{g}_i + u_i - \tilde{u}_i)\right\} \\
&= S^{-1}_{W-\tilde{W}} \sqrt{n} S_{W-\tilde{W}, g-\tilde{g}+u-\tilde{u}}
\end{aligned} \tag{15.72}$$

其中 $g_i = Z_i'\beta(X_i)$。

对于定理的第一部分，我们将证明以下结果：(i) $S_{W-\tilde{W}} = \Phi + o_p(1)$，(ii) $S_{W-\tilde{W}, g-\tilde{g}} = o_p(n^{-1/2})$，(iii) $S_{W-\tilde{W}, \tilde{u}} = o_p(n^{-1/2})$，以及(iv) $\sqrt{n}S_{W-\tilde{W}, \tilde{u}} \to N(0, \Omega)$ 依分布收敛。

(i)的证明：对于矩阵 A 和标量数列 a_n，$A = O_p(a_n)$（$o_p(a_n)$）说明 A 里的每一个元素都是 $O_p(a_n)$（$o_p(a_n)$）。使用式(15.71)，可得

$$S_{W-\widetilde{W}} = S_{\eta+v+(m-\widetilde{m})-\widetilde{v}-\widetilde{\eta}}$$
$$= S_{\eta+v} + S_{(m-\widetilde{m})-\widetilde{v}-\widetilde{\eta}} + 2S_{\eta+v,(m-\widetilde{m})-\widetilde{v}-\widetilde{\eta}}$$

第一项，$S_{\eta+v} = \frac{1}{n}\sum_i(\eta_i+v_i)(\eta_i+v_i)' = \frac{1}{n}\sum_i \varepsilon_i\varepsilon_i' = \Phi + o_p(1)$ 可从大数定律得出。

第二项，$S_{(m-\widetilde{m})-\widetilde{v}-\widetilde{\eta}} \leqslant 3(\bar{S}_{(m-\widetilde{m})} + \bar{S}_{\widetilde{v}} + \bar{S}_{\widetilde{\eta}}) = o_p(1)$ 可从引理 15.10 和引理 15.11(i)得出，并且(iii)将在这一节的最后给出并证明。

最后一项，

$$S_{\eta+v,(m-\widetilde{m})-\widetilde{v}-\widetilde{\eta}} \leqslant \{\bar{S}_{\eta+v}\bar{S}_{(m-\widetilde{m})-\widetilde{v}-\widetilde{\eta}}\}^{1/2}$$
$$= (O_p(1)o_p(1))^{1/2} = o_p(1)$$

由之前的结论得到，其中对于 $m\times m$ 矩阵 A，$\mathrm{diag}(A)$ 是一个 $m\times 1$ 矩阵，由 A 的对角线元素构成，并且对 A 内的每个元素开根号所得的 $A^{1/2}$ 与 A 有同样的维度。

(ii)的证明：利用式(15.71)，可得

$$S_{W-\widetilde{W},g-\widetilde{g}} = S_{\eta+v+(m-\widetilde{m})-\widetilde{v}-\widetilde{\eta},g-\widetilde{g}}$$
$$= S_{\eta+v,g-\widetilde{g}} + S_{m-\widetilde{m},g-\widetilde{g}} - S_{\widetilde{v},g-\widetilde{g}} - S_{\widetilde{\eta},g-\widetilde{g}}$$

对于第一项，注意 η_i+v_i 正交于变系数函数空间$\mathcal{G}$，并且注意 $g_i-\widetilde{g}_i$ 属于$\mathcal{G}$，我们使用引理 15.10，得到

490

$$E[\|S_{\eta+v,g-\widetilde{g}}\|^2] = n^{-2}\sum_{i=1}^n E[(\eta_i+v_i)(\eta_i+v_i)'(g_i-\widetilde{g}_i)^2]$$
$$\leqslant Cn^{-1}\left(\sum_{l=1}^d k_l^{-2\delta_l}\right)E[\|\eta_1+v_1\|^2]$$
$$= O\left(n^{-1}\left(\sum_{l=1}^d k_l^{-2\delta_l}\right)\right) = o(n^{-1})$$

这意味着 $S_{\eta+v,g-\widetilde{g}} = O_p\left(n^{-1/2}\left(\sum_{l=1}^d k_l^{-\delta_l}\right)\right)$。

第二项，由引理 15.10 得

$$S_{m-\widetilde{m},g-\widetilde{g}} \leqslant (\bar{S}_{m-\widetilde{m}}\bar{S}_{g-\widetilde{g}})^{1/2} = O_p\left(\sum_{l=1}^r k_l^{-2\delta_l}\right)$$

第三项，由引理 15.10 和引理 15.11(i)可得

$$S_{\widetilde{v},g-\widetilde{g}} \leqslant (S_{\widetilde{v}}S_{g-\widetilde{g}})^{1/2} = O_p(K/n)O_p\left(\sum_{l=1}^r k_l^{-\delta_l}\right)$$

最后一项，由引理 15.10 和引理 15.11(iii)可得

$$S_{\widetilde{\eta},g-\widetilde{g}} \leqslant (S_{\eta\sim}S_{g-\widetilde{g}})^{1/2} = O_p(K/n)O_p\left(\sum_{l=1}^r k_l^{-\delta_l}\right)$$

把上面的四项结合起来，由假定 15.9 可得

$$S_{W-\tilde{W},g-\tilde{g}} = O_p(n^{-1/2} + (K/n)^{1/2})\left(\sum_{l=1}^{d} k_l^{-\delta_l} + \sum_{l=1}^{d} k_l^{-2\delta_l}\right) = o_p(n^{-1/2})$$

(iii) 的证明：利用式(15.71)，可得

$$S_{W-\tilde{W},\tilde{u}} = S_{\eta+v+(m-\tilde{m})-\tilde{v}-\tilde{\eta},\tilde{u}}$$
$$= S_{\eta+v,\tilde{u}} + S_{m-\tilde{m},\tilde{u}} - S_{\tilde{v},\tilde{u}} - S_{\tilde{\eta},\tilde{u}}$$

第一项，由引理 15.11(ii) 可得

$$S_{\eta+v,\tilde{u}} \leqslant (\bar{S}_{\eta+v}\bar{S}_{\tilde{u}})^{1/2} = O_p(K/n)$$

第二项（使用 tr(AB) = tr(BA)），由引理 15.10 和引理 15.11(ii) 可得

$$S_{m-\tilde{m},\tilde{u}} \leqslant (\bar{S}_{(m-\tilde{m}}\bar{S}_{\tilde{u}})^{1/2} = O_p\left(\sum_{l=1}^{r} k_l^{-\delta_l}\right)O_p(\sqrt{K}/\sqrt{n})$$

第三项，由引理 15.10(i) 和 (ii) 可得

$$S_{\tilde{v},\tilde{u}} \leqslant (\bar{S}_{\tilde{v}}\bar{S}_{\tilde{u}})^{1/2} = O_p(K/n)$$

最后一项，由引理 15.11(ii) 和 (iii) 可得

$$S_{\tilde{\eta},\tilde{u}} \leqslant (S_{\tilde{\eta}}\bar{S}_{\tilde{u}})^{1/2} = O_p(K/n)$$

把上面的四项结合起来，由假定 15.9 可得

$$S_{W-\tilde{W},\tilde{u}} = O_p\left(K/n + n^{-1/2}\sum_{l=1}^{d} k_l^{-\delta_l}\right) = o_p(n^{-1/2})$$ □

491

(iv) 的证明：利用式(15.71)，可得

$$\sqrt{n}S_{W-\tilde{W},u} = \sqrt{n}S_{\eta+v+(m-\tilde{m})-\tilde{v}-\tilde{\eta},u}$$
$$= \sqrt{n}S_{\eta+v,u} + \sqrt{n}(S_{m-\tilde{m},u} - S_{\tilde{v},u} - S_{\tilde{\eta},u})$$

第一项，由 Lindeberg-Feller 中心极限定理，可得

$$\sqrt{n}S_{\eta+v,u} = \sqrt{n}\sum_{i=1}^{n}(\eta_i + v_i)u_i = \sqrt{n}\sum_{i=1}^{n}\varepsilon_i u_i \to N(0,\Omega)$$

依分布收敛。

第二项，由引理 15.10 可得

$$E[S^2_{(m-\tilde{m}),u} \mid X,Z] = \frac{1}{n^2}\mathrm{tr}\{(m-\tilde{m})(m-\tilde{m})'E[uu' \mid X,Z]\}$$
$$\leqslant (C/n)\mathrm{tr}\{(m-\tilde{m})'(m-\tilde{m})/n\}$$
$$= (C/n)S_{m-\tilde{m}} = O_p(n^{-1})O_p\left(\sum_{l=1}^{r} K_l^{-2\delta_l}\right)$$

因此，$S_{m-\tilde{m},u} = O\left(n^{-1/2}\left(\sum_{l=1}^{r} K_l^{-\delta_l}\right)\right)$。

第三项，由引理 15.11(i)可得

$$E[S^2_{\tilde{v},u} \mid X,Z] = \frac{1}{n^2}\mathrm{tr}\{P(P'P)^{-1}P'vv'P(P'P)^{-1}P'E[uu' \mid X,Z]\}$$
$$\leqslant (C/n^2)\mathrm{tr}[P(P'P)^{-1}P'vv'P(P'P)^{-1}P']$$
$$= (C/n)\mathrm{tr}[\tilde{v}\tilde{v}'/n]$$
$$= (C/n)S_{\tilde{v}} = o_p(n^{-1})$$

因此，$S_{\tilde{v},u} = o_p(n^{-1/2})$。 □

最后一项，$S_{\tilde{\eta},u} = o_p(n^{-1/2})$可以用与证明 $S_{\tilde{v},u} = o_p(n^{-1/2})$同样的方法来证明，只是它是引用引理 15.11(iii)，而非引理 15.11(i)。

综合(i)、(ii)、(iii)和(iv)的证明与式(15.72)，我们总结出$\sqrt{n}(\hat{r}-\gamma)\to N(0,\Phi^{-1}\Omega\Phi^{-1})$依分布收敛。

对于定理的第二部分，我们需要证明 $\hat{\Sigma}=\Sigma+o_p(1)$，其中 $\hat{\Sigma}=\hat{\Phi}^{-1}\hat{\Omega}\hat{\Phi}^{-1}$。但是 $\hat{\Phi}=S_{W-\tilde{W}}=\Phi+o_p(1)$已经在上面对(i)的证明中被证明了。利用相同的论证法，容易证明 $\hat{\Omega}=\Omega+o_p(1)$。因此，$\hat{\Sigma}=\Sigma+o_p(1)$。

引理 15.9 $\|\tilde{\alpha}_f-\alpha_f\| = O_p\left(\sum_{l=1}^{r}k_l^{-\delta_l}\right)$，其中 $\tilde{\alpha}_f=(P'P)^{-1}P'f$，$\alpha_f$ 满足假定 15.9，其中 $f=G$ 或 $f=m$。

引理 15.10 $S_{f-\tilde{f}} = O_p\left(\sum_{l=1}^{r}k_l^{-2\delta_l}\right)$，其中 $f=G$ 或 $f=m$。

492

引理 15.11 (i) $S_{\tilde{v}} = O_p(K/n)$，(ii) $S_{\tilde{u}} = O_p(K/n)$，(iii) $S_{\tilde{\eta}} = O_p(K/n)$。

对引理 15.9、15.10 和 15.11 的证明以及引理 15.5、15.7 和 15.8 的证明是相似的，所以我们在这里省略。

15.6.4 定理 15.9 的证明

(i)的证明：注意

$$\hat{u}_i = Y_i - m(X_i,\hat{\gamma}) = u_i - [m(X_i,\hat{\gamma}) - m(X_i,\gamma_0)]$$

我们得到

$$\bar{I}^a_{se} = n^{-1}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}u_i p^K(X_i)'(P'P)^{-1}p^K(X_j)u_j$$
$$-2n^{-1}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}[m(X_i,\hat{\gamma}) - m(X_i,\gamma_0)]$$
$$\times p^K(X_i)'(P'P)^{-1}p^K(X_j)u_j$$

$$+ n^{-1}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}[m(X_i,\hat{\gamma})-m(X_i,\gamma_0)]$$
$$\times p^K(X_i)'(P'P)^{-1}p^K(X_j)$$
$$\times[m(X_i,\hat{\gamma})-m(X_i,\gamma_0)]$$
$$\equiv I_{1n}-2I_{2n}+I_{3n}$$

正如 Newey(1997)所述，我们可以证明 $\hat{I}_{jn}(j=1,2,3)$、$\hat{S}_n$、S_n 对于 $p^K(\boldsymbol{x})$ 的非奇异线性转换都是不变的，于是，我们始终假定 $B=I_k$，即如假定 15.2 定义的 $p^K(x)=P^K(x)$，以及 $E[p^K(\boldsymbol{x})p^K(x)']=I_K$（$I_K$ 是 K 维的单位矩阵），于是 $P'P/n$ 的最小特征值依概率收敛于 1，并且

$$\sup_{1\leqslant i\leqslant n}p^K(X_i)'(P'P)^{-1}p^K(X_i)$$
$$\leqslant\lambda_{\min}^{-1}(P'P/n)\sup_{1\leqslant i\leqslant n}p^K(X_i)'p^K(X_i)/n$$
$$\leqslant\lambda_{\min}^{-1}(P'P/n)\zeta_0^2(K)/n\xrightarrow{p}0$$

其中 $\lambda_{\min}$是 $P'P$ 的最小特征值。应用与 Hong 和 White(1995)对于定理 A.3 的证明类似的推理，我们可得到

$$nI_{1n}/S_n\xrightarrow{d}N(0,1)$$

现在，我们证明 I_{2n}和 I_{3n}都是 $O_p(n^{-1})$阶的。通过使用一个 Talor 级数展开式，可得

493

$$I_{2n}=n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{\gamma}-\gamma_0)'\nabla_\gamma m(X_i,\gamma_0)p^K(X_i)'(P'P)^{-1}p^K(X_j)u_j$$
$$-n^{-1}\sum_{i=1}^{n}(\hat{\gamma}-\gamma_0)'\nabla_\gamma m(X_i,\gamma_0)p^K(X_i)'(P'P)^{-1}p^K(X_i)u_i$$
$$+(1/2)n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{\gamma}-\gamma_0)'\nabla_\gamma^2 m(X_i,\tilde{\gamma})(\hat{\gamma}-\gamma_0)$$
$$\times p^K(X_i)'(P'P)^{-1}p^K(X_j)u_j$$
$$-(1/2)n^{-1}\sum_{i=1}^{n}(\hat{\gamma}-\gamma_0)'\nabla_\gamma^2 m(X_i,\tilde{\gamma})(\hat{\gamma}-\gamma_0)$$
$$\times p^K(X_i)'(P'P)^{-1}p^K(X_i)u_i$$
$$\equiv(\hat{\gamma}-\gamma_0)'(I_{2n,1}-I_{2n,2})+(1/2)(\hat{\gamma}-\gamma_0)'(I_{2n,3}-I_{2n,4})(\hat{\gamma}-\gamma_0)$$

其中 $\tilde{\gamma}$ 是 $\hat{\gamma}$ 和 γ_0 连线上的一点。鉴于 $P(P'P)^{-1}P'$是等幂的，可得

$$E(\|I_{2n,1}\|^2)\leqslant\sum_{k=1}^{K}Cn^{-2}E\Big\{\sum_{i=1}^{n}\sum_{i'=1}^{n}\frac{\partial m(X_i,\gamma_0)}{\partial\gamma_k}$$

$$\times p^{K}(X_i)'(P'P)^{-1}p^{K}(X_i')\frac{\partial m(X_i',\gamma_0)}{\partial\gamma_k}\Big\}$$

$$\leqslant \sum_{k=1}^{K} Cn^{-2}E\left\{\sum_{i=1}^{n}\left[\frac{\partial m(X_i,\gamma_0)}{\partial\gamma_k}\right]^2\right\}$$

$$= Cn^{-2}\sum_{i=1}^{n}E\{\|\nabla m(X_i,\gamma_0)\|^2\} = O(n^{-1})$$

$$E(\|I_{2n,2}\|^2)\leqslant \sum_{k=1}^{K} Cn^{-2}E\Big\{\sum_{i=1}^{n}\left[\frac{\partial m(X_i,\gamma_0)}{\partial\gamma_k}\right]^2$$

$$\times[\sup_i p^{K}(X_i)'(P'P)^{-1}p^{K}(X_i)]^2\Big\}$$

$$= Cn^{-2}\sum_{i=1}^{n}E\{\|\nabla m(x_i,\gamma_0)\|^2\}o_p(1) = o(n^{-1})$$

494 由 Chebyshev 不等式可知 $\|I_{2n,1}-I_{2n,2}\|\leqslant O_p(n^{-1/2})$。对于 $I_{2n,3}$ 和 $I_{2n,4}$、柯西不等式以及 $P(P'P)^{-1}P'$ 是等幂的事实导出

$$\|I_{2n,3}\|^2 = \sum_{k=1}^{K}\sum_{l=1}^{K}\left(n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2 m(X_i,\tilde{\gamma})}{\partial\gamma_k\partial\gamma_l}p^{K}(X_i)'(P'P)^{-1}p^{K}(X_j)u_j\right)^2$$

$$\leqslant \sum_{k=1}^{K}\sum_{l=1}^{K}\left(n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{\partial^2 m(X_i,\tilde{\gamma})}{\partial\gamma_k\partial\gamma_l}p^{K}(X_i)'(P'P)^{-1}p^{K}(X_j)\frac{\partial^2 m(X_i,\tilde{\gamma})}{\partial\gamma_k\partial\gamma_l}\right)$$

$$\times\left(n^{-1}\sum_{j=1}^{n}u_j^2\right)$$

$$\leqslant \sum_{k=1}^{K}\sum_{l=1}^{K}\left(n^{-1}\sum_{i=1}^{n}\left[\frac{\partial^2 m(X_i,\tilde{\gamma})}{\partial\gamma_k\partial\gamma_l}\right]^2\right)\left(n^{-1}\sum_{i=1}^{n}(u_i^2-\sigma_i^2)+n^{-1}\sum_{i=1}^{n}\sigma_i^2\right)$$

$$= O_p(1)$$

其中最后的不等式来自 Chebyshev 不等式。于是,可得

$$\|I_{2n,4}\|^2 = \sum_{k=1}^{K}\sum_{l=1}^{K}\left(n^{-1}\sum_{i=1}^{n}\frac{\partial^2 m(X_i,\tilde{\gamma})}{\partial\gamma_k\partial\gamma_l}p^{K}(X_i)'(P'P)^{-1}p^{K}(X_i)u_i\right)^2$$

$$\leqslant \sum_{k=1}^{K}\sum_{l=1}^{K}\left(n^{-1}\sum_{i=1}^{n}\left[\frac{\partial^2 m(X_i,\tilde{\gamma})}{\partial\gamma_k\partial\gamma_l}\right]^2\right)$$

$$\times\left(n^{-1}\sum_{j=1}^{n}u_j^2\right)[\sup_i p^{K}(X_i)'(P'P)^{-1}p^{K}(X_i)]^2$$

$$= o_p(1)$$

所以，$\| I_{2n,3}-I_{2n,4} \| \leqslant O_p(1)$。于是，

$$|I_{2n}| \leqslant \|\hat{\gamma}-\gamma_0\| \, \|I_{2n,1}-I_{2n,2}\| + (1/2)\|\hat{\gamma}-\gamma_0\|^2 \|I_{2n,3}-I_{2n,4}\|$$
$$\leqslant O_p(n^{-1})$$

通过在均值展开，

$$m(X_i,\hat{\gamma}-m(X_i,\gamma_0)) = \nabla_\gamma m(X_i,\bar{\gamma})'(\hat{\gamma}-\gamma_0), \quad i=1,\cdots,n$$

其中 $\bar{\gamma}$ 是 $\hat{\gamma}$ 和 γ_0 连线上的一点,因此 $\|\tilde{\gamma}-\gamma_0\| \leqslant \|\bar{\gamma}-\gamma_0\|$,我们可以写出

495

$$\begin{aligned}
I_{3n} = {} & n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}(\hat{\gamma}-\gamma_0)'\nabla_\gamma m(X_i,\bar{\gamma})p^K(x_i)'(P'P)^{-1}p^K(X_j) \\
& \times \nabla m(X_i,\bar{\gamma})'(\hat{\gamma}-\gamma_0) \\
& - n^{-1}\sum_{i=1}^{n}(\hat{\gamma}-\gamma_0)'\nabla m(X_i,\bar{\gamma})p^K(x_i)'(P'P)^{-1}p^K(X_j) \\
& \times \nabla m(x_i,\bar{\gamma})'(\hat{\gamma}-\gamma_0) \\
= {} & (\hat{\gamma}-\gamma_0)'(I_{3n,1}-I_{3n,2})(\hat{\gamma}-\gamma_0)
\end{aligned}$$

同样,由柯西不等式和 $P(P'P)^{-1}P'$是等幂的事实,可得

$$\|I_{3n,1}\|^2 \leqslant \sum_{k=1}^{K}\sum_{l=1}^{K}\left(n^{-1}\sum_{i=1}^{n}\left[\frac{\partial m(x_i,\bar{\gamma})}{\partial\gamma_k}\right]^2\right)\left(n^{-1}\sum_{i=1}^{n}\left[\frac{\partial m(x_i,\bar{\gamma})}{\partial\gamma_l}\right]^2\right)$$
$$= O_p(1)$$

且

$$\begin{aligned}
\|I_{3n,2}\|^2 \leqslant {} & [\sup_i p^K(X_i)'(P'P)^{-1}p^K(X_i)]^2 \\
& \times \sum_{k=1}^{K}\sum_{l=1}^{K}\left(n^{-1}\sum_{i=1}^{n}\left[\frac{\partial^2 m(X_i,\bar{\gamma})}{\partial\gamma_k}\right]^2\right)\left(n^{-1}\sum_{i=1}^{n}\left[\frac{\partial^2 m(X_i,\bar{\gamma})}{\partial\gamma_l}\right]^2\right) \\
= {} & o_p(1)
\end{aligned}$$

所以,我们得到 $\|I_{3n,1}-I_{3n,2}\| = O_p(1)$。因此

$$|I_{3n}| \leqslant \|\hat{\gamma}-\gamma_0\|^2 \|I_{3n,1}-I_{3n,2}\| = O_p(n^{-1})$$

概括上面的结果并且使用在 H_0 下的 $S_n = O_p(K^{1/2})$ 和 $\hat{S}_n = S_n + o_p(S_n)$,我们完成了(i)的证明,即

$$\bar{J}^a_{se} = n\bar{I}^a_{se}/S_n + o_p(1) \xrightarrow{d} N(0,1) \qquad \square$$

证明(ii)的证明:在 H_1^a 下,

496

$$\begin{aligned}
\hat{u}_i &= Y_i - m(X_i,\hat{\gamma}) \\
&= u_i - [m(X_i,\hat{\gamma}) - m(X_i,\gamma^*)] - [m(X_i,\gamma^*) - \theta_0(X_i)]
\end{aligned}$$

其中 γ^* 表示 $\hat{\gamma}$ 的概率极限。因为 $\hat{\gamma}-\gamma^* = o_p(1)$ 且 $E(u_i|X_i)=0$,容易证明与

$m(X_i,\hat{\gamma})-m(X_i,\gamma^*)$和$u_i$相关的各项都是$o_p(1)$阶的。于是,在$H_1^a$下,首项是接下来的既不包含$u_i$也不包含$m(X_i,\hat{\gamma})-m(X_i,\gamma^*)$的项。于是,可得

$$\bar{I}_{se}^a=A_n+o_p(1) \tag{15.73}$$

其中

$$A_n=n^{-1}\sum_{i=1}\sum_{j\neq i}^n[m(X_i,\gamma^*)-\theta_0(X_i)]p^K(X_i)'(P'P)^{-1}p^K(X_j)$$
$$\times[m(X_j,\gamma^*)-\theta_0(X_j)]$$

因为存在一个$K\times 1$向量β_K,使得当$K\to\infty$时,

$$E[\{m(\boldsymbol{x},\gamma^*)-\theta_0(\boldsymbol{x})-\boldsymbol{p}^K(\boldsymbol{x})'\beta_K\}^2]\to 0$$

通过加减项,可得

$$\sum_{j\neq i}^n p^K(X_i)'(P'P)^{-1}p^K(X_j)[m(X_j,\gamma^*)-\theta_0(X_j)]$$
$$=[m(X_i,\gamma^*)-\theta_0(X_i)]+[p^K(X_i)'\beta_K-m(X_i,\gamma^*)+\theta_0(X_i)]$$
$$+\sum_{j\neq i}^n p^K(X_i)'(P'P)^{-1}p^K(X_j)[m(X_j,\gamma^*)-\theta_0(X_j)-p^K(X_i)'\beta_K] \tag{15.74}$$

将式(15.74)代入式(15.73),注意与$[p^K(fX_i)'\beta_K-m(X_i,\gamma^*)+\theta_0(X_i)]$相关的项都是$o_p(1)$阶的,可得

$$A_n=n^{-1}\sum_i[m(X_i,\gamma^*)-\theta_0(X_i)]^2$$
$$+n^{-1}\sum_i[m(X_i,\gamma^*)-\theta_0(X_i)]$$
$$\times[p^K(X_i)'\beta_K-m(X_i,\gamma^*)+\theta_0(X_i)]$$
$$=n^{-1}\sum_i\sum_j[m(X_i,\gamma^*)-\theta_0(x_i)]p^K(X_i)'(P'P)^{-1}p^K(X_j)$$
$$\times[m(X_j,\gamma^*)-\theta_0(X_j)-p^K(X_j)'\beta_K] \tag{15.75}$$

497 因为$E\{[p^K(x)'\beta_K-m(x,\gamma^*)+\theta_0(x)]^2\}=o_p(1)$,可得

$$\bar{I}_{se}^a\xrightarrow{p}E\{[m(X_i,\gamma^*)-\theta_0(X_i)]^2\}\equiv C>0$$

因此,由式(15.75)以及$S_n^*=O_p(K^{1/2})$和$S_n=S_n^*+o_p(S_n^*)$的事实,我们得到在H_1^a下,$\bar{J}_{se}^a=n\bar{I}_{se}/\hat{S}_n=O_p(n/K^{1/2})$。(ii)也立即得证。 □

15.6.5 定理15.10的证明

在这一节我们使用没有下标i的字母(一般为大写)来表示向量或者矩阵。

比如说，$\mathcal{H}(X,x)$、U、$\hat{U}$、m 和 $\phi(x)$ 都是 $n\times 1$ 向量，其第 i 个元素分别是$\mathcal{H}(X_i,x)$、u_i、$\hat{u}_i$、$m(X_i)$ 和 $\phi_i(x)$。当然对于 $n\times 1$(或 $d\times 1$)向量$\mathcal{A}$，我们用 $\|\mathcal{A}\|$ 去表示它的欧几里得范数。

定理15.10(i)的证明：注意

$$\hat{u}_i = Y_i - z_0(X_i)'\hat{\gamma} - p^K(X_i)'\hat{\beta}$$
$$= u_i - z_0(X_i)'(\hat{\gamma}-\gamma) + m(X_i) - \hat{m}(X_i)$$

且 $\hat{m}(X_i) \equiv p^K(X_i)'\hat{\beta} = p^K(X_i)'(P'P)^- P'(Y - Z_0\hat{\gamma})$。

于是我们可得，用向量—矩阵表示法有

$$\hat{m} = P(P'P)^- P'(Y - Z_0\hat{\gamma}) = M_n(Y - Z_0\hat{\gamma})$$
$$= M_n[U - Z_0(\hat{\gamma}-\gamma) + m]$$
$$\hat{U} = U - M_nU - (I_n - M_n)Z_0(\hat{\gamma}-\gamma) + (I_n - M_n)m \tag{15.76}$$

利用式(15.76)可得

$$\hat{J}_n(x) = n^{-1/2}\sum_i \mathcal{H}(X_i,x)\hat{u}_i = n^{-1/2}(\mathcal{H}(X,x))'\hat{U}$$
$$= n^{-12}(\mathcal{H}(X,x))'U - n^{-1/2}(\mathcal{H}(X,x)'MnU$$
$$- n^{-1/2}(\mathcal{H}(X,x))'(I_n - M_n)Z_0(\hat{\gamma}-\gamma)$$
$$+ n^{-1/2}(\mathcal{H}(X,x))'(I_n - M_n)m$$
$$\equiv J_{n1}(x) - J_{n2}(x) - J_{n3}(x) + J_{n4}(x) \tag{15.77}$$

引理15.12证明了 $\|J_{n2}(\cdot) - n^{-1/2}\phi(\cdot)'U\|_\nu = o_p(1)$，其中 $\phi(x)$ 是一个 $n\times 1$ 向量，其第 i 个分量是 $\phi_i(x) = E_\zeta[\mathcal{H}(X_i,x)]$。引理15.14建立了 $\|J_{n4}(\cdot) - n^{-1/2}\psi(\cdot)'U\|_\nu = o_p(1)$，其中 $\psi(x)$ 是一个 $n\times 1$ 向量，其第 i 个分量是 $\psi_i(x) = E[\mathcal{H}(X_i,x)\varepsilon'_i]\{E[\varepsilon_i\varepsilon'_i]\}^{-1}\varepsilon_i$，且 $\varepsilon_i = z_0(X_i) - E_\mathcal{G}[z_0(X_i)]$。 498

引理15.13证明了 $\|J_{n3}(\cdot)\|_\nu = o_p(1)$。

定义

$$J_n(x) = n^{-1/2}\sum_i[\mathcal{H}(X_i,x) - \phi_i(x) - \psi_i(x)]u_i$$
$$\equiv n^{-1/2}\sum_i Z_i(x)$$

然后，由引理15.12到15.14，可得

$$\|\hat{J}_n(\cdot) - J_n(\cdot)\|_\nu = o_p(1) \tag{15.78}$$

容易看出 $E[\|J_n(\cdot)\|_\nu^2] < \infty$，即 $J_n(\cdot)$ 是紧凑的。于是，通过 Hilbert 值随机数组的中心极限定理可得

$$J_n(\cdot)\text{ 在}\mathcal{L}_2(\mathcal{S},\nu,\|.\|_\nu)\text{ 上弱收敛于 }J_\infty(\cdot) \tag{15.79}$$

其中 $J_\infty(\cdot)$ 是一个高斯过程，其具有零均值且其协方差函数为

$$\begin{aligned}\Sigma(x,x') &= \mathrm{cov}(J_n(x),J_n(x')) = E[Z_i(x)Z_i(x')] \\ &= E\{\sigma^2(X_i)[\mathcal{H}(X_i,x) - \phi_i(x) - \psi_i(x)] \\ &\quad \times [\mathcal{H}(X_i,x') - \phi_i(x') - \psi_i(x')]\}\end{aligned}$$

等式(15.78)意味着 $\hat{J}_n(\cdot)$ 和 $J_n(\cdot)$ 有同样的极限分布，于是，它和式(15.79)都意味着 $\hat{J}_n(\cdot)$ 弱收敛于 $J_\infty(\cdot)$。于是我们完成了定理 15.10(i)的证明。 □

定理 15.10(ii) 的证明：很显然 $h(J) \stackrel{\text{def}}{=} \int [J(x)]^2 F(dx)$ 是在 $\mathcal{L}_2(\mathcal{S},F)$ 上的一个连续函数。给定 F 关于 Lebesgue 测度 ν 是完全连续的，$h(J)$ 在 $\mathcal{L}_2(\mathcal{S},\nu)$ 上也是连续的。因此，通过定理 15.10(i) 和连续映射定理，可得 $\int [\hat{J}_n(x)]^2 F(dx)$ 依分布收敛于 $\int [J_\infty(x)]^2 F(dx)$。

现在，定义 $A_n = \mathrm{CM}_n - h(\hat{J}_n^2)$。Li, Hsiao 和 Zinn(2003)证明了 $A_n = o_p(1)$。因此，由定理 15.10(i)和连续映射定理的结论可得

$$\begin{aligned}\mathrm{CM}_n &= \int [\hat{J}_n(x)]^2 F(dx) + A_n = \int [\hat{J}_n(x)]^2 F(dx) + o_p(1) \\ &\to \int [J_\infty(x)]^2 F(dx) \text{ 依分布收敛}\end{aligned}$$

499 这完成了定理 15.10(ii)的证明。

下面我们给出在定理 15.10 的证明中使用到的一些引理。对于一个 $n \times d$ 矩阵 A，我们记 $\tilde{A} = M_n A$ 并且令 $\tilde{A}_i$ 为 $\tilde{A}$ 的第 i 个分量。例如，$\tilde{m} = M_n m$，$\tilde{Z}_0 = M_n Z_0$。

引理 15.12 $\| J_{n2}(\cdot) - n^{-1/2}\phi(\cdot)'U \|_\nu^2 = o_p(1)$，其中 $\phi(x)$ 是一个 $n \times 1$ 向量，其第 i 个元素是 $\phi_i(x) = E_{\mathcal{G}}[\mathcal{H}(X_i,x)]$。

证明：定义 $V_i(x) = \mathcal{H}(X_i,x) - \phi_i(x)$。那么 $E_{\mathcal{G}}[V_i(x)] = 0$ 并且 $E_{\mathcal{G}}[V_i^2(x)]$ 对于任意的 $x \in \mathcal{S}$ 是有界的，同时我们由引理 15.15 和 15.16 可得

$$\begin{aligned}&E[\| J_{n2}(\cdot) - n^{-1/2}\phi(\cdot)'U \|_\nu^2 \mid X] \\ &= n^{-1}\int [(\mathcal{H}(X,x)'M_n) - \phi(x)']E(UU' \mid X) \\ &\quad \times [M_n\mathcal{H}(X,x) - \phi(x)]\nu(dx) \\ &\leqslant Cn^{-1}\int [M_n\mathcal{H}(X,x) - \phi(x)]'[M_n\mathcal{H}(X,x) - \phi(x)]\nu(dx) \\ &= Cn^{-1}\| M_n\mathcal{H}(X,x) - \phi(x) \|_\nu^2 \\ &= Cn^{-1}\| M_n(\phi(x) + V(x)) - \phi(x) \|_\nu^2\end{aligned}$$

$$\leqslant 2Cn^{-1}\{\|M_n\phi(x)-\phi(x)\|_\nu^2+\|M_nV(x)\|_\nu^2\}$$

$$=O_p\left(\sum_{l=1}^{L}K_l^{-2\delta_l}+K/n\right)=o_p(1)$$

引理 15.13 $\|J_{n3}(\cdot)-n^{-1/2}\psi(\cdot)U\|_\nu^2=o_p(1)$。

证明:注意 $z_0(X_i)-\tilde{Z}_0(X_i)$ 估计了 $\varepsilon_i=z_0(X_i)-E_{\mathcal{G}}[z_0(X_i)]$,或者以矩阵表示 $Z_0-M_nZ_0$ 估计了 ε。由引理 15.17 可知$(\hat{\gamma}-\gamma)=\{E[\varepsilon_i\varepsilon_i']\}^{-1}n^{-1}\sum_i\varepsilon_iu_i+o_p(n^{-1/2})$。使用引理 15.15 和 15.16 可得

$$\|n^{-1}\mathcal{H}(X_i,\cdot)'(I_n-M_n)Z_0-E[\mathcal{H}(X_i,\cdot)'\varepsilon_i]\|_\nu^2=o_p(1)$$

因此,

$$\begin{aligned}J_{n3}(\cdot)&=n^{-1/2}\mathcal{H}(X_i,\cdot)'(I_n-M_n)Z_0(\hat{\gamma}-\gamma)\\&=E[\mathcal{H}(X_i,\cdot)\varepsilon_i][n^{-1/2}(\hat{\gamma}-\gamma)]+o_p(1)\\&=E[\mathcal{H}(X_i,\cdot)\varepsilon_i]\{E[\varepsilon_i\varepsilon_i']\}^{-1}\left[n^{-1/2}\sum_i\varepsilon_iu_i\right]+o_p(1)\\&=n^{-1/2}\psi(\cdot)U+o_p(1)\end{aligned}$$

□

引理 15.14 $\|J_{n4}(\cdot)\|_\nu^2=o_p(1)$。 500

证明:由引理 15.15,

$$\begin{aligned}\|J_{n4}(\cdot)\|_\nu^2&\leqslant n^{-1}\sum_i\sum_j\int\mathcal{H}(X_i,x)\,\mathcal{H}(X_j,x)(m_i-\tilde{m}_i)(m_j-\tilde{m}_j)\nu(dx)\\&\leqslant C\sum_i\int[(m_i-\tilde{m}_i)^2]\nu(dx)\\&=C[\|m-\tilde{m}\|_\nu^2]=nO_p\left(\sum_{l=1}^{L}K_l^{-2\delta_l}\right)=o_p(1)\end{aligned}$$

引理 15.15 令 $f_i(x)\equiv f_0(x,X_i)\in\mathcal{G}$(可加函数类),其中 $f_0(x,X_i)$ 是 $d\times1$ 维的(d 是一个有限的正整数)。令 $f_X(x)$ 表示 $n\times d$ 矩阵,其第 i 行是 $f_i(x)'$。定义 $\tilde{f}_X(x)=M_nf_X(x)$,那么

$$n^{-1}\|f_X(x)-M_nf_X(x)\|_\nu^2=O_p\left(\sum_{l=1}K_l^{-2\delta_l}\right)=o_p(1)$$

证明:由 Andrew 和 Whang(1990)以及 Newey(1995,1997)的结论,或者由 Li(2000)的引理 A.4,

$$\begin{aligned}n^{-1}E[\|f_X(x)-M_nf_X(x)\|_\nu^2]&\equiv n^{-1}E[\|f_X(x)-\tilde{f}_X(x)\|_\nu^2]\\&=n^{-1}\int E[\|f_X(x)-\tilde{f}_X(x)\|_\nu^2]\nu(dx)\\&=O_p\left(\sum_lK_l^{-2\delta_l}\right)\end{aligned}$$

引理 15.16 令 $v_i(x) \equiv V(x, X_i)$，$E_{\mathcal{G}}(v_i(x)) = 0$ 且 $E_{\mathcal{G}}([v_i(x)]^2)$ 在 $x \in \mathcal{S}$ 上是一致有界的。同样，定义 $V(x) = (v_1(x), \cdots, v_n(x))'$，以及 $\tilde{V}(x) = M_n V(x)$。然后，可得

$$n^{-1} \| M_n V(\cdot) \|_\nu^2 = n^{-1} \| \tilde{V}(\cdot) \|_\nu^2 = O_p(K/n) = o_p(1)$$

501 证明：不失一般性，我们可以假定 $E[p^K(X_i) p^K(X_i)'] = I_K$（见 15.6 节开始部分的论证，在定理 15.1 的证明之前）。首先我们证明

$$E[\| P'V(\cdot)/n \|_\nu^2] = O((K/n)^{1/2})$$

注意 $p^K(X_i) \in \mathcal{G}$和 $v_i(\cdot) \perp \mathcal{G}$意味着 $E[p^K(X_i) \nu_i(\cdot)] = 0$。可得

$$\begin{aligned}
&E[\| P'V(\cdot)/n \|_\nu^2] \\
&= n^{-2}\Big\{ \sum_i \int E[v_i(x)^2 p^K(X_i)' p^K(X_i)] \nu(dx) \\
&\quad + \sum_i \sum_{j \neq i} \int E[v_i(x) p^K(X_i)'] \times E[v_j(x) p^K(X_j)] \nu(dx) \Big\} \\
&= n^{-1} \int E[v_1(x)^2 p^K(X_1)' p^K(X_1)] \nu(dx) \\
&\leqslant C n^{-1} E[p^K(X_1)' p^K(X_1)] = O(K/n)
\end{aligned}$$

这意味着

$$\| P'V(x)/n \|_\nu^2 = O_p(K/n) = o_p(1) \tag{15.80}$$

那么由式(15.80)和 $\| (P'P/n)^- - I \| = O_p(\zeta_0(K)\sqrt{K}/\sqrt{n}) = o_p(1)$ 的事实（见 Newey(1997, pp. 161—162)关于定理 1 的证明），

$$\begin{aligned}
n^{-1} \| M_n V(\cdot) \|_\nu^2 &= n^{-1} \int V(x)' M_n V(x) \nu(dx) \\
&= \int (V(x)'P/n) (P'P/n)^- (P'V(x)/n) \nu(dx) \\
&= \int (V(x)'P/n) [I + (P'P/n)^- - I] (P'V(x)/n) \nu(dx) \\
&= \int \| P'V(x)/n \|^2 [1 + o_p(1)] \nu(dx) \\
&= \int O_p(K/n) [1 + o_p(1)] \nu(dx) \\
&= O_p(K/n) = o_p(1)
\end{aligned}$$

引理 15.17

$$(\hat{\gamma} - \gamma) = \{E[\varepsilon_i \varepsilon_i']\}^{-1} \Big\{ n^{-1} \sum_i \varepsilon_i u_i \Big\} + o_p(n^{-1/2})$$

502 其中 $\varepsilon_i = z_0(X_i) - E_{\mathcal{G}}[z_0(X_i)]$。

这在 Li(2000)的定理 2.1 中已经证明。注意引理 15.17 意味着 $\hat{\gamma}-\gamma=O_p(n^{-1/2})$。

15.7　习　题

习题 15.1　假定 $E(u_i^2|X_i)=E(u_i^2)=\sigma^2$,证明定理 15.2。

提示:定义 $A_n=n^{-1/2}V_k^{-1/2}(\hat{g}(x)-g(x))$。然后证明(i),由假定 15.2,

$$E[A_n]=O((nK)^{1/2}K^{-\alpha})=o(1)$$

$$\mathrm{var}[A_n]=nV_k^{-1}\mathrm{var}(\hat{g}(x))=nV_K^{-1}[n^{-1}V_K+o(1)]=1+o(1)$$

可证实 Liapunov 的中心极限定理条件对 A_n 成立。于是,$A_n\xrightarrow{d}N(0,1)$。进一步,证明(ii) $\hat{V}_K=V_K+o_p(1)$。

定理 15.2 由(i)和(ii)可得。

习题 15.2　推导式(15.14)。

下面的三个习题参考 15.3.2 节(非线性可加成分的选择)。

习题 15.3　对使用幂级数估计量的可加模型编写一个程序,并且使用 Huang 和 Yang(2004)的 BIC 标准选择显著性变量。

习题 15.4　重复习题 15.3,但是使用去一法选择序列项的数目。

习题 15.5　证明在式(15.35)中定义的 $\theta_{\alpha l}$是由 $\theta_{lj}=\int_{-1}^{1}g_\alpha(z_\alpha)p_l(z_\alpha)\mathrm{d}z_\alpha$ 给出。

提示:对式(15.35)的两边都乘以 $p_s(z_\alpha)$,关于 z_α 积分并且利用

$$\int p_l(z_\alpha)p_s(z_\alpha)\mathrm{d}z_\alpha=\delta_{ls}$$

习题 15.6　证明式(15.60)。

提示:(i) 定义

$$I_{n,se}^{a,0}=n^{-2}\sum_i\sum_{j\neq i}u_ip^K(x_i)'(P'P)^{-}p^K(x_j)u_j$$

$$=2n^{-2}\sum_i\sum_{j>i}u_ip^K(x_i)'(P'P)^{-}p^K(x_j)u_j$$

计算 $E[(n/K^{1/2})I_{n,se}^{a,0}]$和 $\mathrm{var}[(n/K)^{1/2}I_{n,se}^{a,0}]$,并且证明在 H_0^a 下,

$$(n/K)^{1/2}I_{n,se}^{a,0}/\hat{\sigma}_{se,a}^2\xrightarrow{d}N(0,1)$$

(ii) 证明 $I_{n,se}^a-I_{n,se}^{a,0}=o_p(K^{1/2}/n)$。等式(15.60)可以由(i)和(ii)得出。

第 5 部分
时间序列、联立方程和面板数据模型

第16章　工具变量和半参数模型的有效估计

505

回归元和随机干扰项在参数回归模型中是不相关的这个假定(即回归元是外生的)对于推导一致性来说是一个关键的假定。甚至当其他的经典假定都成立时,如果这个假定不成立,那么最小二乘法就失去了其作为一个估计量的吸引力。在很多社会科学应用中,这个假定可能会被违背。一个经常用于这种情况的替代方法被称为"工具变量"方法或者简称为"IV 估计"。然而,参数模型并不是唯一的一类当回归元是内生时遭受影响的模型。在这一章,我们考虑一组为了处理内生回归元的出现而被发展的半参数估计量。

16.1　在参数部分具有内生回归元的部分线性模型

我们首先考虑一个下列形式的半参数部分线性时间序列模型[①]

$$Y_t = X_t'\beta + \theta(Z_t) + u_t, \quad t = 1, \cdots, n \tag{16.1}$$

506

其中 X_t 是 $p\times 1$ 维,β 是一个 $p\times 1$ 维的未知参数向量,Z_t 的维度是 q,并且 $\theta(\cdot)$ 是一个非指定的平滑函数。我们允许 X_t 包含 Y_t 的滞后值;比如 X_t 的第一个元素可能是 Y_{t-1},因此,式(16.1)构建了一个非参数动态模型。我们将假定 Z_t 在 $E(u_t|Z_s)=0, s\leqslant t$ 的意义上来说是弱外生性的。对依赖数据(比如时间序列)的非参数估计的渐

① 见第18章对依赖数据非参数估计的渐近理论。

近理论在第 18 章讨论。

我们允许误差 u_t 是序列相关的这种可能。如果 $\theta(z)=z'\gamma$ 关于 z 是线性的，式(16.1)成为一个参数线性模型，我们可以选择一个 Z_{t-1} 的线性组合，比如 $Z'_{t-1}\alpha$，作为关于 Y_{t-1} 的一个工具变量（α 是一个 $q\times 1$ 常数向量），因为根据我们 Z_t 是弱外生性的假定，可以证明在一般情况下 $E(Y_{t-1}Z'_{t-1}\alpha)\neq 0$，以及 $E(u_tZ'_{t-1}\alpha)=0$。于是，如果模型关于 Z_t 是线性的，那么 Z'_{t-1} 是 Y_{t-1} 一个适当的工具变量。然而，对于一般函数 $\theta(\cdot)$，Y_{t-1} 和 Z_{t-1} 有可能是不相关的。例如，为了阐释的简单，假定 Z_t 是一个标量，并且假定 $\theta(Z_t)=Z_t^2$，数据是严格平稳的，且 $E(Z_t^3)=0$。然后，可以证明 $E(Y_{t-1}Z_{t-1})=0$（习题 16.1）。因此，当 $\theta(\cdot)$ 的函数形式未知时，我们不能把 Z_{t-1} 作为 Y_{t-1} 的工具来使用。由式(16.1)，Y_{t-1} 与 Z_{t-1} 的某些函数显然是相关的，比如 $\theta(Z_{t-1})$。如果我们限制自己根据 Z_{t-1} 的函数去构建一个工具，那么，根据 Newey(1990a)，可以证明在这种情形下的最优工具是 $E(Y_{t-1}|Z_{t-1})$。这个最优工具函数是未知的，但是可以用任一非参数方法，比如核方法、k-nn 方法或者序列方法，一致地估计出来。

等式(16.1)包含一个未知函数 $\theta(\cdot)$。由 Robinson(1988)，我们首先消除 $\theta(\cdot)$。对式(16.1)关于 Z_t 取期望，然后，由式(16.1)减去它，可得

$$Y_t-E(Y_t\mid Z_t)=(X_t-E(X_t\mid Z_t))'\beta+u_t\equiv v_t'\beta+u_t \tag{16.2}$$

其中我们使用 $E(u_t|Z_t)=0$ 和 $v_t\stackrel{\text{def}}{=}X_t-E(X_t|Z_t)$。注意 v_t 和 u_t 可能是相关的，因为 v_t 包含 Y_t 的滞后值，并且 u_t 可能是序列相关的。假定存在一个 $d\times 1$ $(d\geqslant p)$ 工具变量 W_t，其与 X_t 相关但是与 u_t 不相关（详见下面的式(16.5)）。把式(16.2)改写为向量—矩阵形式，X 和 v $(v=X-E(X|Z))$ 都是 $n\times p$ 维，W 是 $n\times d$ 维，Y、$E(Y|Z)$ 和 u 都是 $n\times 1$ 维。将式(16.2)的矩阵形式乘以 W'，可得

507

$$W'(Y-E(Y\mid Z))=W'v\beta+W'u \tag{16.3}$$

对式(16.3)使用最小二乘法，由下式可得 β 的 IV-OLS 估计量

$$\begin{aligned}\tilde{\beta}_{\text{IVO}}&=(v'WW'v)^{-1}v'WW'(Y-E(Y\mid Z))\\&=\beta+(v'WW'v)^{-1}v'WW'u\end{aligned} \tag{16.4}$$

其中 W 和 v 分别是 $n\times d$ 和 $n\times p$ 维，其典型的行由 W_t'和 v_t'给出。

我们引入下列条件来推导 $\tilde{\beta}_{\text{IVO}}$的渐近分布：

$$W'v/n\xrightarrow{p}E[W_tv'_t]\equiv A$$

$$\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{n}E[W_tW'_su_tu_s]\xrightarrow{p}B$$

$$W'u/\sqrt{n}\xrightarrow{d}N(0,B) \tag{16.5}$$

其中我们要求 A 是列满秩的并且 B 是一个有限的正定矩阵。注意当数据关于 t 是独立同分布时(如横截面数据),或者 u_t 是一个鞅差分过程时,那么在式(16.5)中 $B=E[\sigma^2(W_t)W_tW'_t]$($\sigma^2(W_t)=E(u_t^2|W_t)$)。对于一般的时间序列数据,当 u_t 不是一个鞅差分过程时,B 的表达式依赖于 u_t 的特定误差过程。

利用式(16.5),容易证明

$$\sqrt{n}(\tilde{\beta}_{\text{IVO}}-\beta)\to N(0,V_O)\text{ 依分布收敛} \tag{16.6}$$

其中 $V_O=Q^{-1}A'BAQ^{-1}$,$Q=A'A$,其中 A 和 B 由式(16.5)定义。式(16.6)的证明留作习题(见习题16.2)。

接下来我们考虑一种情形,当 u_t 是鞅差分过程并且是条件同方差时,即误差项 u_t 使得 $E(u_t|W_t,Z_t,\Omega_{t-1})=0$($\Omega_\tau$ 表示一直到时期 τ 的信息集),并且 $E(u_t^2|W_t,Z_t,\Omega_{t-1})=E(u_t^2)=\sigma^2$。式(16.3)中的误差 $W'u/\sqrt{n}$ 的方差为 $B=\sigma^2E[W_tW'_t]$。在这种情形下,我们可以使用下面的IV-GLS估计量估计 β:

$$\tilde{\beta}_{\text{IVG}}=(v'W(W'W)^{-1}W'_v)^{-1}v'W(W'W)^{-1}W'(Y-E(Y|Z)) \tag{16.7}$$

508

为了推导 $\tilde{\beta}_{\text{IVG}}$ 的渐近分布,我们做出下面的假定

$$W'v/n\xrightarrow{p}E[W_tv'_t]\equiv A$$

$$W'W/n\xrightarrow{p}E[W_tW'_t]=\bar{B}$$

$$W'u/n\xrightarrow{p}E[W_tu_t]=0$$

$$W'u/\sqrt{n}\xrightarrow{d}N(0,\sigma^2\bar{B}) \tag{16.8}$$

如果我们使用式(16.8),可以直接证明

$$\sqrt{n}(\tilde{\beta}_{\text{IVG}}-\beta)\to N(0,V_G)\text{ 依分布收敛} \tag{16.9}$$

其中 $V_G=\sigma^2(A\bar{B}^{-1}A')^{-1}$,其中 A 和 $\bar{B}$ 的定义见式(16.8)。

可以证明 V_G-V_O 是负半定的(见White(1984))。因此,$\tilde{\beta}_{\text{IVG}}$ 是比 $\tilde{\beta}_{\text{IVO}}$ 更加渐近有效的。当条件同方差误差的假定被违背时,$\tilde{\beta}_{\text{IVG}}$ 仍然是一个对于 β 的 $\sqrt{n}$ 一致估计量,但是它可能有一个不同的渐近方差,并且可能不会比 $\hat{\beta}_{\text{IVO}}$ 更加有效。

记 $X'_t=(Y_{t-1},\cdots,Y_{t-d_1},X'_{t2})$,其中 X_{t2} 的维度是 $d-d_1$,假定 X_{t2} 是弱外生性的,即对所有 $t\leqslant s$,$E(u_s|X_{t2})=0$。当我们想使用一个 Z_{t-s} 的函数为 Y_{t-s} 构建一个工具时,我们就需要为 $X'_{t1}=(Y_{t-1},\cdots,Y_{t-d_1})$ 寻找工具。如上文讨论的,$W_{ts}=E(Y_{t-s}|Z_{t-s})$ 应该被选为 Y_{t-s} 的工具。令 $\hat{W}_t=(\hat{W}_{t1},\cdots,\hat{W}_{td_1})$,其中 $\hat{W}_{ts}=\hat{E}(Y_{t-s}|Z_{t-s})$,$s=1,\cdots,d_1$。令 $\hat{v}_t=X_t-\hat{E}(X_t|Z_t)$ 和 $\hat{Y}_t=\hat{E}(Y_t|Z_t)$ 分别表示

$X_t - E(X_t|Z_t)$和$E(Y_t|Z_t)$的非参数估计量，那么，可行的 IV - GLS 估计量为

$$\hat{\beta}_{\mathrm{IVO}} = (\hat{v}'\hat{W}\hat{W}'\hat{v})^{-1}\hat{v}'\hat{W}\hat{W}'(Y - \hat{E}(Y|Z)) \tag{16.10}$$

如果误差是条件同方差的，我们可用可行的 IV-GLS 估计量估计β：

$$\hat{\beta}_{\mathrm{IVG}} = (\hat{v}'\hat{W}(\hat{W}'\hat{W})^{-1}\hat{W}'\hat{v})^{-1}\hat{v}'\hat{W}(\hat{W}'\hat{W})^{-1}\hat{W}'(Y - \hat{E}(Y|Z)) \tag{16.11}$$

在非参数估计均方误差有比$n^{-1/2}$更小阶的条件下，可以证明：

$$\sqrt{n}(\hat{\beta}_{\mathrm{IVO}} - \beta) \rightarrow N(0, V_O) \text{ 依分布收敛} \tag{16.12}$$

509 其中V_O是在式(16.6)中被定义的，如果误差u_t是鞅差分过程并且是条件同方差的，那么

$$\sqrt{n}(\hat{\beta}_{\mathrm{IVG}} - \beta) \rightarrow N(0, V_G) \text{ 依分布收敛} \tag{16.13}$$

其中V_G与式(16.9)的定义是一样的。

如果我们使用非参数核方法，式(16.12)和式(16.13)的证明是与第7章中定理7.1的证明类似的。令$\tilde{\beta}$表示$\tilde{\beta}_{\mathrm{IVO}}$或者$\tilde{\beta}_{\mathrm{IVG}}$，$\hat{\beta}$表示$\hat{\beta}_{\mathrm{IVO}}$或者$\hat{\beta}_{\mathrm{IVG}}$。可以记$\sqrt{n}(\hat{\beta}-\beta) = \sqrt{n}(\hat{\beta}-\tilde{\beta}) + \sqrt{n}(\tilde{\beta}-\beta)$。$\hat{\beta}-\tilde{\beta}$是$\left(\sum_{s=1}^{q} h_s^{2\nu} + (nh_1\cdots h_q)^{-1}\right)$阶的($\nu$是核函数的阶数)。如果$\hat{\beta}-\tilde{\beta}$是$o_p(n^{-1/2})$阶的，那么我们由式(16.6)和式(16.13)可得$\sqrt{n}(\hat{\beta}-\beta) = \sqrt{n}(\tilde{\beta}-\beta) + o_p(1) \rightarrow N(0, V)$。于是，可行估计量$\hat{\beta}$与不可行估计量$\tilde{\beta}$具有相同的(一阶)渐近分布。

16.2 在参数部分具有内生回归元的一个变系数模型

我们考虑一个具有下列形式的横截面或者时间序列数据的变系数模型

$$Y_t = X_t'\beta(Z_t) + U_t,\ t = 1, \cdots, n \tag{16.14}$$

其中X_t是$p \times 1$维的，β是未知函数的$p \times 1$向量，并且Z_t是q维的。我们允许X_t包含内生变量，比如$E(u_t|X_t) \neq 0$。我们假定Z_t是(弱)外生的，即$E(u_t|Z_s) = 0, s \leq t$，如果它是一个时间序列，或者可以简化为$E(u_t|Z_t) = 0$，如果我们有独立的数据(如横截面数据)。

假定存在一个工具变量向量v_t使得$E(u_t|v_t) = 0$，且假定$E[E(X_t|v_t)X_t']$是非奇异的。那么，我们可以使用$W_t = E(X_t|v_t)$作为X_t的一个工具。例如，对于时间序列数据，如果$X_t = Y_{t-1}$且Z_t是外生的，那么我们可以使用选择$v_t = Z_{t-1}$，这可以得出$W_t = E(Y_{t-1}|Z_{t-1})$。

令W和X为$n \times p$矩阵，Y和u为$n \times 1$矩阵，以及$K_{h,z}$为一个$n \times n$对角矩

510 阵，其第t个对角元素为$K_{h,tz} = \prod_{s=1}^{q} h_s^{-1} k((Z_{ts} - Z_s)/h_s)$，那么

$$\begin{aligned}\tilde{\beta}(z) &= (W'K_{h,z}X)^{-1}W'K_{h,z}Y \\ &= \beta(z) + (W'K_{h,z}X)^{-1}W'K_{h,z}u \\ &= \beta(z) + \left[\sum_{t=1}^{n} W_t X'_t K_{h,tz}\right]^{-1}\left[\sum_{t=1}^{n} W_t u_t K_{h,tz}\right] \end{aligned} \tag{16.15}$$

在数据都是独立同分布的假定下，关于 $\beta(z)$ 和 $f(x,\omega,z)$ 的某些平滑性假定成立，并且关于 (x_t,ω_t,z_t,u_t) 的某些矩条件成立，然后我们在 16.4 节证明

$$\sqrt{nh_1\cdots h_q}\left[\tilde{\beta}(z) - \beta(z) - \sum_{s=1}^{q} h_s^2 B_s(z)\right] \xrightarrow{d} N(0,\Omega_z) \tag{16.16}$$

其中

$$\begin{aligned} B_s(z) &= (1/2)\kappa_2 M_z^{-1}\{f_z(z)\beta_{ss}(z)E[W_tX_t' \mid Z_t = z] \\ &\quad + 2\beta_s(z)E[W_tX_t' f_s(X_t,W_t,Z_t)f(X_t,W_t \mid Z_t = z)^{-1} \mid Z_t = z|]\} \\ M_z &= f_z(z)E[W_tX_t' \mid Z_t = z] \\ \Omega_Z &= M_Z^{-1}V_ZM_Z^{-1} \\ V_z &= \kappa^q f_z(z)E[W_tX_t'\sigma^2(W_t,X_t,Z_t) \mid Z_t = z] \end{aligned}$$

$$\sigma^2(W_t,X_t,z) = E[u_t^2 \mid W_t,X_t,Z_t = z]$$

注意如果 X_t 是外生的，那么我们可以选择 $W_t = X_t$，并且式(16.16)简化为第 9 章中考虑的结果。

式(16.16)中给出的正态结论对于弱依赖混合数据仍然是可行的，给定混合系数满足一定的衰变率并且误差项 u_t 是鞅差分过程。

最优工具函数 $W_t = E[X_t \mid v_t]$ 是未知的，但是可以用任一非参数方法，比如核方法、k-nn 或序列方法一致估计出来。如果使用核方法，那么，对于 $\beta(z)$ 一个可行的估计量可以通过在式(16.15)中把 W_t 用 $\hat{W}_t = \hat{E}(X_t \mid v_t) = \sum_{j=1}^{n} X_jK((v_j - v_t)/b)/\sum_{j=1}^{n} K((v_j - v_t)/b)$ 替代得到。令 $\hat{\beta}(z)$ 表示 $\beta(z)$ 最终的可行估计量，然后，在与 Cai, Das, Xiong 和 Wu(2006), Cai 和 Li(2005)，以及 Das(2005) 给出的相似的正则条件下，比如 $\sum_{s=1}^{q} b_s^2 = o\left(\sum_{s=1}^{q} h_s^2\right)$，可以证明 $\hat{\beta}(z)$ 与式(16.16)中给出的 $\beta(z)$ 具有同样的渐近分布。

在上面的半参数模型中，我们仅仅考虑了参数部分是内生的情形，然而，非参数部分仍然是外生的。当非参数部分是内生的，渐近分析将会更加复杂。在接下来的一节中我们讨论一般的半参数模型，即在非参数部分中可能出现内生变量。在第 17 章我们将在一个完全的非参数体系中讨论内生回归元。 511

16.3 Ai 和 Chen 的有条件矩限制的有效估计量

Ai 和 Chen(2003)为了在下面形式的条件矩限制下估计一个计量模型考虑了一个一般的框架

$$E[\rho(Z,\theta_0,g_0(\cdot))\mid X]=0 \tag{16.17}$$

其中误差 $\rho(\cdot)$ 是一个已知函数形式的标量[①], θ_0 是一个 $d\times 1$ 的有限维参数向量(模型的参数部分),并且 $g_0(\cdot)=(g_{01},g_{02},\cdots,g_{0q})$ 是未知函数向量。含有未知函数 $g_0(\cdot)$ 允许式(16.17)包括许多重要类型的半参数和非参数模型。比如说,它包括在第 7 章中研究的作为特殊情形的部分线性模型, $\rho(Z,\theta_0,g_0)=Y-X_1'\theta_0-g_0(X_2)$,以及在第 8 章中研究的单一指数模型, $\rho(Z,\theta_0,g_0)=Y-g_0(X'\theta_0)$。

在 Ai 和 Chen(2003)中考虑的一个主要的例子是这样一种部分线性模型,内生变量进入模型的非参数部分,即 $\rho(Z,\theta_0,g_0(\cdot))=Y_1-X_1'\theta_0-g_0(Y_2)$,对应的回归模型是 $Y_1=X_1'\theta_0+g_0(Y_2)+u$,但是因为 Y_2 是内生的, $E(u\mid X_1,Y_2)\neq 0$。在这个情形下,我们需要假定存在一个工具变量向量 X_2,使得 $E(u\mid X_1,X_2)\equiv E[Y_1-X_1'\theta_0-g_0(Y_2)\mid X_1,X_2]=0$($X_2$ 作为 Y_2 的一个工具)。

16.3.1 估计过程

当 Z 包含内生变量时,我们不能通过标准最小距离或者将广义矩估计法(以下简称 GMM)应用于 ρ 估计出 $\alpha_0\overset{\text{def}}{=}(\theta_0,g_0)$。在式(16.17)识别出 $\alpha_0=(\theta_0,g_0)$ 的假定下,Ai 和 Chen(2003)建议,首先把 $\rho(Z,\alpha)$ 投射到外生变量 X 的空间上。定义 $m(x,\alpha)=E[\rho(Z,\alpha)\mid X=x]$,然后,用最小距离估计量估计 $\alpha=(\theta,g)$,即最小化 $E[m(X,\alpha)^2/\sigma^2(X)]$,其中 $\sigma^2(X)=\text{var}(\rho(Z,\alpha_0)\mid X)$。注意,在实践中, $m(\cdot)$ 是未知的原因有二:(i) $g_0(\cdot)$ 有一个未知的函数形式,(ii)即使我们知道 g_0,给定 X 则 ρ 的条件期望仍然是未知的。

512

如果 g 是已知的,我们可以用非参数的方法估计 $m(\cdot)$,即序列方法。令 $p_j(X)(j=1,2,\cdots)$ 表示已知的基本函数序列(幂级数、样条、傅里叶级数等),并且令 $p^{K_n}(X)=(p_1(X),\cdots,p_{K_n}(X)')$,那么 $m(x,\alpha)$ 的非参数序列估计量如下:

$$\hat{m}(X,\alpha)=p^{K_n}(X)'\hat{\pi} \tag{16.18}$$

其中 $\hat{\pi}=(P'P)^{-1}P'\rho(\cdot,\alpha)$, $P=(p^{K_n}(X_1),\cdots,p^{K_n}(X_n))'$ 是 $n\times K_n$ 维的,并且 $\rho(\cdot,\alpha)$ 是 $n\times 1$ 维的,其第 i 行是 $\rho(Z_i,\alpha)$。用 $m(X,\alpha)$ 的非参数序列估计量

① Ai 和 Chen(2003)考虑了这样一般的情形,其中 ρ 是一个向量值函数。这里,为了表达的简洁性,我们只考虑 ρ 是一个标量的情形。

$\hat{m}(X,\alpha)$，如果 g 是已知的，我们可以通过最小化下式估计 θ：

$$\frac{1}{n}\sum_{i=1}^{n}[\hat{m}(X_i,\theta,g)]^2/\hat{\sigma}^2(X_i)$$

其中 $\hat{\sigma}^2(X)$ 是 $\sigma^2(X)$ 的一致估计，引入 $\sigma^2(X)$ 是为了处理潜在的异方差性。

然而，事实上 g 也是未知的。Ai 和 Chen(2003)建议使用 B 样条、傅里叶级数或者小波过滤基(wavelet sieve basis)逼近 $g(Z)$。令 $q^{L_n}(\cdot)=(q_1(\cdot),\cdots,q_{L_n}(\cdot))'$ 表示前 L_n 个基础函数，那么，可以使用 $q^{L_n}(Z)'\beta$ 逼近 $g(Z)$，其中 β 在假定 $\theta_0\in\Theta$ 下是 $L_n\times 1$ 维的，Θ 是 $\mathbb{R}^d$ 的一个紧子集。这里，假定基础函数 $q^{L_n}\in\mathcal{H}_n$，其中 $\mathcal{H}_n$ 是一个可以计算的有限维度紧参数空间的过滤空间，其在连续函数空间内随着 n 增大变得密集。

在式(16.18)中对一个 K_n 的给定值，(θ,g) 的估计成为 (θ,β) 的估计。可以用 $(\hat{\theta},\hat{\beta})$ 估计 (θ,β)，它是由下面的最小化问题的解来定义：

$$\begin{bmatrix}\hat{\theta}\\ \hat{\beta}\end{bmatrix}=\arg\min_{(\theta,\beta)\in\Theta\times\mathbb{R}^{L_n}} n^{-1}\sum_{i=1}^{n}\hat{m}(X_i,\theta,q^{L_n}(Z_i)'\beta)^2/\hat{\sigma}^2(X_i)\qquad(16.19)$$

513

$g(z)$ 的最终估计量是 $\hat{g}(z)=q^{L_n}(z)'\hat{\beta}$。

加权函数 $\sigma^2(X)$ 是用于有效性的考虑。如果误差是条件同方差的，我们可以简单地用 1 代替 σ^2，或者，如果主要对得到 (θ,g) 的一致估计量感兴趣，但是不太在意(渐近)有效性，也可以用 1 替代 $\sigma^2(X)$。在这种情形下，可以简单地最小化(关于 (θ,β))下式：

$$n^{-1}\sum_{i=1}^{n}[\hat{m}(X_i,\theta,q^{L_n}(Z_i)'\beta)]^2$$

上面的半参数估计量可以被理解为 GMM 估计量。令 $(\tilde{\theta},\tilde{\beta})$ 表示最终的估计量，结合式(16.18)和式(16.19)且用 1 替代 $\hat{\sigma}^2(X)$，可得

$$\begin{bmatrix}\tilde{\theta}\\ \tilde{\beta}\end{bmatrix}=\min_{(\theta,\beta)\in\Theta\times\mathbb{R}^{L_n}}\left(\sum_{j=1}^{n}\rho(Z_j,\theta,q^{L_n}(Z_j)'\beta)p^{K_n}(X_j)\right)'(P'P)^{-1}$$
$$\times\left(\sum_{j=1}^{n}\rho(Z_j,\theta,q^{L_n}(Z_j)'\beta)p^{K_n}(X_j)\right)\qquad(16.20)$$

我们将在下一节讨论 $\hat{\theta}$ 的渐近分布。

16.3.2　$\hat{\theta}$ 的渐近正态性

假定对于几乎所有的 Z，$\rho(Z,(1-\tau)\alpha_0+\tau\alpha_0)$ 在 $\tau=0$ 是连续可微的。表示在 α_0 处在 $[\alpha-\alpha_0]$ 方向的一阶方向导数为

$$\frac{\mathrm{d}\rho(Z,\alpha_0)}{\mathrm{d}\alpha}[\alpha-\alpha_0] \overset{\mathrm{def}}{=} \frac{\mathrm{d}\rho(Z,(1-\tau)\alpha_0+\tau\alpha_0)}{\mathrm{d}\tau}\Big|_{\tau=0}$$

且定义

$$\frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}\alpha}[\alpha_1-\alpha_2] \overset{\mathrm{def}}{=} E\left\{\frac{\mathrm{d}\rho(Z,\alpha_0)}{\mathrm{d}\alpha}[\alpha_1-\alpha_2]\,|\,X\right\}$$

令$\mathcal{A}=\mathbb{R}^d\times\mathcal{W}$表示$\alpha=(\theta,g)$的可行范围,其中$\mathcal{W}$是排除$g_0$的 Hölder 连续类函数;进一步细节见 Ai 和 Chen(2003)。对于任何$\alpha_1,\alpha_2\in\mathcal{A}$,我们有

$$\frac{\mathrm{d}\rho(Z,\alpha_0)}{\mathrm{d}\alpha}[\alpha_1-\alpha_2] = \frac{\mathrm{d}\rho(Z,\alpha_0)}{\mathrm{d}\alpha}[\alpha_1-\alpha_0] + \frac{\mathrm{d}\rho(Z,\alpha_0)}{\mathrm{d}\alpha}[\alpha_2-\alpha_0]$$

514 为了研究$\hat{\alpha}$收敛于α的速率,必须定义一个测度(metric),一般使用的包括 sup 和L_2测度。Ai 和 Chen(2003)定义了一个以下的弱测度$\|\cdot\|_w$:

$$\|\alpha_1-\alpha_2\|_w = \sqrt{E\left\{\left[\frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}\alpha}[\alpha_1-\alpha_2]\right]^2/\sigma^2(X)\right\}}$$

并且L_2测度是(在$\|\cdot\|_w$中变$m(X,\cdot)$为$\rho(Z,\cdot)$)

$$\|\alpha_1-\alpha_2\|_2 = \sqrt{E\left\{\left[\frac{\mathrm{d}\rho(Z,\alpha_0)}{\mathrm{d}\alpha}[\alpha_1-\alpha_2]\right]^2/\sigma^2(X)\right\}}$$

可以证明对所有$\alpha\in A$,$\|\alpha-\alpha_0\|_w\leqslant\|\alpha-\alpha_0\|_2$。

当未知函数g依赖于内生变量Y时,在平常的L_2测度下我们有缓慢的收敛速度,但是在$\|\cdot\|_w$下有较快的收敛速度。Ai 和 Chen(2003)证明对半参数估计量$\hat{\theta}_n$的$\sqrt{n}$正态性,它满足在弱测度$\|\cdot\|_w$下有一个较快的收敛速度。

定义

$$\frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}\alpha}[\alpha-\alpha_0] = \frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}\theta'}[\theta-\alpha_0] + \frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}g}[g-g_0]$$

对于θ_j(属于θ)的每个分量,$j=1,\cdots,d$,令$w_j^*\in\mathcal{W}$表示下面的最优化问题的解:

$$\min_{w_j\in\mathcal{W}} E\left\{\left(\frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}\theta_j} - \frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}g}[w_j]\right)^2/\sigma^2(x)\right\} \tag{16.21}$$

定义$w^*=(w_1^*,\cdots,w_d^*)$,可得

$$\frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}g}[w^*] = \left(\frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}g}[w_1^*],\cdots,\frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}g}[w_d^*]\right)$$

且

$$D(X) \overset{\mathrm{def}}{=} \frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}\theta'} - \frac{\mathrm{d}m(X,\alpha_0)}{\mathrm{d}g}[w^*]$$

Ai 和 Chen(2003)证明了下面的结论。

定理 16.1 在 Ai 和 Chen(2003)给出的正则性条件下,有

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, V_0) \text{ 依分布收敛}$$

其中
$$V_0 = \{E[D(X)D(X)'/\sigma^2(X)]\}^{-1}。$$

Ai 和 Chen(2003)也证明对 θ_0 的任意$\sqrt{n}$一致半参数估计量的渐近方差,V_0 等于半参数有效的下边界。 515

16.3.3 在非参数部分具有内生回归元的部分线性模型

对于具有内生非参数成分的部分线性模型,我们有

$$\rho(Z_i, \alpha) = Y_{1i} - X'_{1i}\theta - g(Y_{2i}),\ E[\rho(Z_i, \alpha_0) \mid X_{1i}, X_{2i}] = 0 \qquad (16.22)$$

其中 $\alpha = (\theta, g)$,Y_1 是一个标量,且 $Z = (Y_1, Y'_2, X'_1)'$,$X = (X'_1, X'_2)'$。$g(\cdot)$的函数形式是未知的,且 Ai 和 Chen(2003)建议使用一个样条—小波过滤基去逼近 $g(Y_2)$。令 $q^{L_n}(\cdot) = (q_1(\cdot), \cdots, q_{L_n}(\cdot))'$表示前 L_n 个基础函数,那么,我们可以使用 $q^{L_n}(Y_2)'\beta$ 去逼近 $g(Y_2)$,其中 β 是 $L_n \times 1$ 维的。由式(16.20)可知,$(\hat{\theta}, \hat{\beta})$最小化下面的目标函数:

$$\rho_\alpha P(P'P)^{-1}P'\rho_\alpha$$

其中 ρ_α 是一个 $n \times 1$ 向量,其第 i 个元素是($\alpha = (\theta, g)$)

$$\rho_{i,\theta,\beta} = Y_{1i} - X'_{1i}\theta - q^{L_n}(Y_{2i})'\beta \qquad (16.23)$$

且 P 是 $n \times K_n$ 维的,其第 i 行元素是 $P^{K_n}(X_i)' = (p_{K_n1}(X_i), \cdots, p_{K_nK_n}(X_i)')$。

容易证明,我们也可以把$(\hat{\theta}, \hat{\beta})$理解为下面的最小化问题的解:

$$\begin{bmatrix} \hat{\theta} \\ \hat{\beta} \end{bmatrix} = \arg \min_{(\theta,\beta) \in \Theta \times \mathbb{R}^{L_n}} \sum_{j=1}^{n} \hat{\rho}^2_{i,\theta,\beta} \qquad (16.24)$$

其中

$$\hat{\rho}_{i,\theta,\beta} = p^{K_n}(X_i)'(P'P)^{-1}P(Y_{1i} - X'_{1i}\theta - q^{L_n}(Y_{2i})'\beta) \qquad (16.25)$$

注意 $\hat{\rho}_{i,\theta,\beta}$是由 $\rho(Z_i, \theta, q^{L_n}(Y_{2i})'\beta) \equiv Y_{1i} - X'_{1i}\theta - q^{L_n}(Y_{2i})'\beta$ 对 $P^{K_n}(X_i)$进行回归所得到的拟合值。

注意式(16.23)(或者等价地式(16.24))导致 $\hat{\theta}$ 和 $\hat{\beta}$ 的一个闭式解,因为 ρ 对于 θ 是线性的,且 g 和 g 的过滤估计量对于 β 是线性的(因为我们使用的是线性过滤)。事实上,具有一个相同权重函数的半参数估计过程是简单地对 $Y_{1i} = X'_{1i}\theta_0 + q^{L_n}(Y_{2i})'\beta + u_i$ 应用二阶段最小二乘法估计,其中 $P^{K_n}(X_i)$是作为 $q^{L_n}(Y_{2i})$的工具。 516

为了推出 $\hat{\theta}_n$ 的渐近分布,注意对于具有相同权重的部分线性模型(即 $m(X, \alpha) = X'_1\theta + E[g(Y_2) \mid X]$),式(16.21)成为

$$\min_{w_j \in \mathcal{W}} E\{(X_{1j} - [w_j(Y_2)])'(X_{1j} - [w_j(Y_2)])\} \tag{16.26}$$

于是,我们得到 $w_j^* = E[w_j(Y_2) | X]$。

对具有内生非参数成分的部分线性模型应用定理 16.1 得到下面的推论。

推论 16.1 令 w^* 为下式的解

$$\min_{w_j} E\{[X_{1j} - E[w_j(Y_2) | X]]^2\}, j = 1, \cdots, d$$

且假定 $E[g_0(Y_2) | X]$不在 X_1 的线性区域①,那么

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, V_0) \text{依分布收敛}$$

其中 $V_0 = \{E[D(X)D(X)'/\sigma^2(X)]\}^{-1}$,且 $D(X) = X_1' - E[w^*(Y_2) | X]$。

Ai 和 Chen(2003)也证明 $\hat{g}(y_2) = q^{L_n}(y_2)'\hat{\beta}$ 是 $g(y_2)$的一个一致估计量,其收敛速度小于 $O_p(n^{-1/4})$,即 $\hat{g}(y_2) - g(y_2) = o_p(n^{-1/4})$。

目前为止,我们关注 Y_2 是内生的。然而,如果 $Y_2 = X_2$ 是外生的,且假定 $\operatorname{var}(Y_1 - X_1'\theta - g(X_2) | X) = \sigma^2$(即条件同方差误差),那么上面的结论变为

$$\begin{aligned} D(x) &= \min_g E\{[X_1 - E(g(X_2) | X)]'[X_1 - E(g(X_2) | X)]\} \\ &= \min_g E\{[X_1 - g(X_2)]'[X_1 - g(X_2)]\} \\ &= E\{[X_1 - E(X_2 | X_1)]'[X_1 - E(X_2 | X_1)]\} \end{aligned}$$

其中第二个等式来自 $E(g(X_2) | X_1, X_2) = g(X_2)$的事实,且最后一个等式来自定理2.1,即 $E(X_1 | X_2)$是作为 X_2 的函数在均方误差意义上的 X_1 的最优预测。于是,在 Y_2 为外生的情形下,可得 $V = \sigma^2\{E[DD']\}^{-1}$, $D = X_1 - E(X_2 | X_1)$,这是在第 7 章讨论的 Robinson(1988)的结论。

517 16.4 公式(16.16)的证明

接下来我们使用 $\|\cdot\|$ 表示欧几里得范数。我们观察到

$$\begin{aligned} \tilde{\beta}(z) &= [\sum_j W_j X'_j K_{h,z_j,z}]^{-1} \sum_j W_j Y_j K_{h,z_j,z} \\ &= [\sum_j W_j X'_j K_{h,z_j,z}]^{-1} \sum_j W_j (X'_j \beta(Z_j) + u_j) K_{h,z_j,z} \\ &= [\sum_j W_j X'_j K_{h,z_j,z}]^{-1} \times \sum_j w_j (X'_j \beta(z) + X'_j(\beta(Z_j) - \beta(z)) + u_j) K_{h,z_j,z} \\ &= \beta(z) + [\sum_j W_j X'_j K_{h,z_j,z}]^{-1} \times \sum_j W_j (X'_j(\beta(Z_j) - \beta(z)) + u_j) K_{h,z_j,z} \\ &= \beta(z) + [D_n(z)]^{-1}\{A_{1n}(z) + A_{2n}(z)\} \end{aligned}$$

① 也就是说,对所有 $\theta \in \mathbb{R}^d$, $E\{[X_1'\theta - g_0(Y_2)]^2\} > 0$。

其中

$$D_n(z) = n^{-1}\sum_j W_j X'_j K_{h,z_j,z}$$

$$\mathcal{A}_{1n}(z) = n^{-1}\sum_j W_j X'_j(\beta(Z_j) - \beta(z))K_{h,z_j,z}$$

$$\mathcal{A}_{2n}(z) = n^{-1}\sum_j W_j u_j K_{h,z_j,z}$$

下面的(i)—(iii)意味着式(16.16)：

(i) $D_n(z) = n^{-1}\sum_j W_j X'_j K_{h,z_j,z} \xrightarrow{p} M_z$

(ii) $\mathcal{A}_{1n}(z) = \sum_{s=1}^{q} h_s^2 B_s(z) + o_p\left(\sum_{s=1}^{q} h_s^2\right)$

(iii) $\sqrt{nh_1\cdots h_q}A_{2n}(z) \xrightarrow{d} N(0, V_z)$

接下来我们证明这些结论。 518

(i) 的证明：令 $\|h\|^2 = \sum_{s=1}^{q} h_s^2$，可得

$$\begin{aligned}
E(D_n(z)) &= E[W_1 W'_1 K_{h,z_1,z}] \\
&= \int q_1 x'_1 K_h(z_1 - z) f(x_1, w_1, z_1)\mathrm{d}x_1\mathrm{d}z_1\mathrm{d}w_1 \\
&= \int w_1 x'_1 K(v) f(x_1, w_1, z + hv)\mathrm{d}v\mathrm{d}x_1\mathrm{d}w_1 \\
&= \left[\int w_1 x'_1 f(x_1, w_1, z)\mathrm{d}x_1\mathrm{d}w_1\right]\left[\int K(v)\mathrm{d}v + O(|h|^2)\right] \\
&= f_z(z)\left[\int w_1 x'_1 f(x_1, w_1 | z_1 = z)\mathrm{d}x_1\mathrm{d}w_1\right][1 + O(|h|^2)] \\
&= f_z(z)E[w_1 x'_1 | z_1 = z] + o(1) = M_z + o(1)
\end{aligned}$$

同样，可以容易证明

$$\mathrm{var}(D_n(z)) = O((nh_1\cdots h_q)^{-1}) = o(1)$$

因此，已经证明了

$$D_n(z) = M_z + o_p(1) \tag{16.27}$$

□

(ii) 的证明： 519

$$\begin{aligned}
E[A_{1n}(z)] &= E[W_j X'_j(\beta(Z_j) - \beta(z))K_{h,z_j,z}] \\
&= \int w_1 x'_1(\beta(z_1) - \beta(z))K_{h,z_1,z} f(x_1, w_1, z_1)\mathrm{d}x_1\mathrm{d}w_1\mathrm{d}z_1 \\
&= \int w_1 x'_1(\beta(z + hv) - \beta(z)) f(x_1, w_1, z + hv)K(v)\mathrm{d}x_1\mathrm{d}w_1\mathrm{d}v
\end{aligned}$$

$$= \int w_1 x'_1 \left[\sum_{s=1}^{q} h_s \beta_s(z) v_s + (1/2) \sum_{s=1}^{q} h_s^2 \beta_{ss}(z) v_s^2 \right]$$

$$\times \left[f(x_1, w_1, z) + \sum_{s=1}^{q} f_s(x_1, w_1, z) h_s v_s \right] K(v) \mathrm{d}x_1 \mathrm{d}w_1 \mathrm{d}v + O\left(\sum_{s=1}^{q} h_s^3 \right)$$

$$= \kappa_2 \sum_{s1}^{q} h_s^2 \int w_1 x'_1 [\beta_s(z) f_s(x_1, w_1, z) / f(x_1, w_1 \mid z_1 = z)$$

$$+ (1/2) f_z(z) \beta_{ss}(z)] f(x_1, w_1 \mid z_1 = z) \mathrm{d}x_1 \mathrm{d}w_1 + O\left(\sum_{s=1}^{q} h_s^3 \right)$$

$$= M_z \sum_{s=1}^{q} h_s^2 B_s(z) + O\left(\sum_{s=1}^{q} h_s^3 \right)$$

其中 M_z 和 $B_s(z)$ 在式(16.16)中被定义。

类似地,可以证明

$$\mathrm{var}(A_{1n}(z)) = O\left(\sum_{s=1}^{q} h_s^2 (nh_1 \cdots h_q)^{-1} + \sum_{s=1}^{q} h_s^5 \right)$$

因此,

$$\left(A_{1n}(z) - \sum_{s=1} h_s^2 B_s(z) \right) = o_p\left(\sum_{s=1}^{q} h_s^2 + (nh_1 \cdots h_q)^{-1/2} \right) \quad (16.28)$$

□

(iii) 的证明: $\sqrt{nH_q} A_{2n}(z)$ 具有零均值且它的方差为($H_q = h_1 \cdots h_q$)

520

$$(nH_q)^{-1} \left\{ \sum_{j=1}^{n} E[W_j W'_j u_j^2 K_{h,z_j,z}^2] \right\}$$

$$= H_q^{-1} \int f(w_1, z_1) \sigma_u^2(w_1, z_1) w_1 x'_1 K_{h,z_1,z}^2 \mathrm{d}w_1 \mathrm{d}z_1$$

$$= \left[\int w_1 w_1{}' \sigma_u^2(w_1, z) f(w_1, z) \mathrm{d}w_1 \right] \left[\int K^2(v) \mathrm{d}v \right] + O\left(\sum_{s=1}^{q} h_s \right)$$

$$= \kappa^q f_z(z) E[W_i W'_i \sigma_u^2(W_i, Z_i) \mid Z_i = z] + o(1) = V_z + o(1)$$

其中 $\kappa^q = \int K^2(v) \mathrm{d}v$。

直接可以验证 Liapunov 中心极限定理的条件成立。于是

$$\sqrt{nH_q} A_{2n}(z) \xrightarrow{d} N(0, V_z) \quad (16.29)$$

□

结合式(16.27)、(16.28)和(16.29),我们证明了

$$\sqrt{nH_q} \left(\hat{\beta}(z) - \beta(z) - \sum_{s=1}^{q} h_s^2 B_s(z) \right)$$

$$= [D_n(z)]^{-1}\sqrt{nH_q}\left\{A_{1n}(z) - M_z\sum_{s=1}^{q}h_s^2B_s(z) + A_{2n}(z)\right\}$$

$$= [M_z + o_p(1)]^{-1}\{\sqrt{nH_q}A_{2n}(z) + o_p(1)\} \xrightarrow{d} M_z^{-1}N(0,V_z)$$

$$= N(0,M_z^{-1}V_zM_z^{-1})$$

16.5 习　题

习题 16.1　假定在式(16.1)中 $\theta(Z_t)=Z_t^2(q=1)$，且 $X_t=Y_{t-1}(p=1)$，假定数据是严格平稳的，在这些条件下证明 $E(Y_{t-1}Z_{t-1})=0$。

习题 16.2　(i)证明式(16.6)，(ii)证明式(16.9)。

习题 16.3　对于 16.3.2 节中定义的 $\|\cdot\|_w$ 和 $\|\cdot\|_2$，证明对所有 $\alpha\in\mathcal{A}$，$\|\alpha-\alpha_0\|_w\leqslant\|\alpha-\alpha_0\|_2$。

第 17 章　非参数回归模型的内生性

521

两类一般的计量经济学模型是单方程和多方程模型。当在一个方程中作为因变量出现的变量也在其他方程中作为解释变量出现时,这一系列方程在本质上被称为"联立"的。在这一章我们考虑在非参数背景下为联立方程建模的各种方法。

17.1　一个非参数模型

我们考虑下面的非参数回归模型,

$$Y_i = g(X_i) + \varepsilon_i, \quad i = 1, \cdots, n \qquad (17.1)$$

其中 $g(\cdot)$ 是未知形式的一个函数。我们的目的是估计 $g(\cdot)$。然而,不像在第 2 章中给出的回归模型,这里的误差项 ε_i 现在被认为与解释变量 X_i 相关,使得 $E(\varepsilon_i | X_i) \neq 0$。所以,条件均值与关注的目标不一致,即 $E(Y_i | X_i) \neq g(X_i)$。正如 Hall 和 Horowitz(2003)所解释的,如果有另一个变量同时对 X_i 和 Y_i 引起变化却不包含在模型中,这种情况就有可能出现。比如,假定 Y_i 表示个人 i 的每小时工资,X_i 包括该个人的教育水平等变量。误差项 ε_i 包括没观察到的个人特征,比如"能力"。高能力的个体很可能会选择较高的教育水平,因此,能力和教育水平是相关的,即误差项 ε_i 至少与 X_i 的某些部分是相关的。在这种情况下,直接应用第 2 章中所给出的估计量将会产生与所要目标不一致的估计。

522

就像在线性参数回归模型中的情形,当存在内生性的回归元时,我们需要假定存在某个工具变量,称为 Z_i,

使得

$$E(\varepsilon_i \mid Z_i) = 0 \tag{17.2}$$

这有时候被称为“工具外生性”,然而我们也需要假定 Z_i 和 X_i 之间有很强的关系,这有时候被称为“工具相关性”。给定 Z_i 的存在和一组规则条件,就有可能获得 $g(\cdot)$ 的一致估计。但是,估计 $g(\cdot)$ 的问题非常复杂,因为从约简式 $E(Y_i|Z_i)$ 到结构式 $g(X_i)$ 的映射证明是不连续的,使得构造一个一致估计量很困难。目前,在式(17.2)成立的假定下,我们将延缓估计式(17.1)的讨论,并且我们应该首先考虑一个简单的三角半参数联立方程模型。

17.2 一个三角联立方程模型

Newey 等(1999)考虑了如下三角非参数联立方程模型:

$$Y_i = g_1(X_i, Z_{1i}) + \varepsilon_i \tag{17.3}$$

$$X_i = \prod(Z_i) + u_i \equiv E(X_i \mid Z_i) + u_i \tag{17.4}$$

其中 Y_i 是一个标量,X_i 和 u_i 是 d_x 维的,Z_{1i}是 d_1 维的,并且是 $Z_i=(Z_{1i},Z_{2i})$ 的一部分,它是 q 维向量。Z_i 被假定为外生的,满足 $E(u_i|Z_i)=0$ 且 $E(\varepsilon_i|u_i,Z_i)=E(\varepsilon_i|u_i)$。也就是说,式(17.4)是约简式方程。$g_1(\cdot)$ 和 $\prod(\cdot)$ 的函数形式没有设定。我们的主要兴趣在于一致的 $g_1(\cdot)$。

我们将令 $g_2(u_i)=E(\varepsilon_i|u_i)$。注意式(17.4)意味着 $E(\cdot|X_i,Z_i)=E(\cdot|u_i,Z_i)$,因为 X_i 是 (Z_i,u_i) 的一个确定性函数,这与式(17.3)共同导致 523

$$\begin{aligned} E(Y_i \mid X_i, Z_i) &= g_1(X_i, Z_{1i}) + E(\varepsilon_i \mid X_i, Z_i) \\ &= g_1(X_i, Z_{1i}) + E(\varepsilon_i \mid u_i) \\ &= g_1(X_i, Z_{1i}) + g_2(u_i) \equiv g(W_i) \end{aligned} \tag{17.5}$$

其中 $W_i=(X_i',Z_i',u_i')'$。方程(17.5)显示 $E(Y_i|X_i,Z_i)$ 是一个可加的回归模型,我们在第 9 章中讨论过,不同之处是这里 u_i 是不可观测的。但是通过非参数核函数或序列方法,从式(17.4)可以得到 u_i 的一致估计。给定序列方法施加给定结构更加方便,Newey 等(1999)建议用序列方法从式(17.5)中估计出 $g_1(\cdot)$,其中 u_i 从式(17.4)中估计。

Newey 等(1999)证明,如果 (x,z_1) 与 u 之间没有函数联系,那么,$g_1(x,z_1)$ 则能被识别为可加常数。这跟在第 9 章中讨论的识别条件是相似的。这就能够满足许多应用,比如,当兴趣在于由 x 和 z_1 的变化而导致的 y 的变化时。但是在其他情况下,比如当要求预测需求数量时,知道 $g_1(x,z_1)$ 的水平是很理想的,如果也假定 $E(\varepsilon)=0$,使得 $E(Y)=E[g_1(X,Z_1)]$,它就可以被识别出来。

更正式地，Newey 等(1999)证明了如下识别条件：

引理 17.1 如果 $g_1(\cdot)$、$g_2(\cdot)$ 和 Π 是可微的，(z_i, u_i) 的支撑集的边界有零概率，并且，以概率 1，$\partial\Pi(z_i)/\partial z_{2i}$ 的秩是 d_x，则 $g_1(x, z_1)$ 是可识别的。

Newey 等(1999)为估计 $g_1(\cdot)$ 提出了一个两阶段估计方法。在第一阶段，令 $r^L(z) = (r_{1L}(z), \cdots, r_{LL}(z))'$ 为一个逼近(基础)函数的向量，并且令 $\hat{\Pi}(z)$ 为 X_i 对 $r_i = r^L(Z_i)$ 回归的预测值，即

$$\hat{\Pi}(z) = r^L(z)'\hat{\gamma}, \hat{\gamma} = (R'R)^{-1}R'(X_1, \cdots, X_n)', R = (r_1, \cdots, r_n)' \quad (17.6)$$

在第二阶段，令 $p^K(w) = (p_{1K}(w), \cdots, p_{KK}(w))'$ 为 $w = (x', z_1', u')'$ 的一个逼近函数向量，其具有一个可加结构，使得每个 $p_{KK}(w)$ 不是依赖于 (x, z_1) 就是依赖于 u，但不是两者都依赖的。也就是说，逼近基本函数与式(17.5)具有同样的可加结构，即在基础函数中 (x, z_1) 和 u 之间没有交互项。令 $\hat{u}_i = X_i - \hat{\Pi}(Z_i)$，$\hat{W}_i = (X_i', Z_{1i}', \hat{u}_i')'$，且令 $\mathbf{1}(\mathcal{A})$ 表示事件$\mathcal{A}$的示性函数，我们引入一个以下形式的修剪函数：

524

$$\gamma(w) = \prod_{j=1}^{d+d_x} \mathbf{1}(a_j \leqslant w_j \leqslant b_j) \quad (17.7)$$

其中 a_j 和 b_j 是有限常数，且 w_j 是 w 的第 j 个分量。对修剪函数的使用仅仅是为了简化渐近理论。第二阶段包括 Y_i 对 $\hat{p}_i = p^K(\hat{W}_i)$ 作回归，此处，由于修剪数据的限制可得

$$\begin{aligned} \hat{g}(w) &= p^K(w)'\hat{\beta} \\ \hat{\beta} &= (\hat{P}'\hat{P})^{-1}\hat{P}'\mathcal{Y} \\ \hat{P} &= (\hat{\tau}_1\hat{p}_1, \cdots, \hat{\tau}_n\hat{p}_n)' \\ \mathcal{Y} &= (Y_1, \cdots, Y_n)' \end{aligned} \quad (17.8)$$

其中 $\hat{\boldsymbol{\tau}}_i = \boldsymbol{\tau}(\hat{W}_i)$，$\hat{W}_i = (X_i', Z_{1i}', \hat{u}_i)'$。

通过把依赖于 (x, z_1) 的项和依赖于 u 的项归并起来，估计量 $\hat{g}(w)$ 可以被用于构造 $g_1(x, z_1)$ 和 $g_2(u)$ 的估计。如果 $p_{1K}(w) = 1$，那么，因为接下来的 K_g 项 $p_{kK}(w)$ 依赖于 (x, z_1)，剩余的项仅仅依赖于 u。于是，$g_1(\cdot)$ 和 $g_2(\cdot)$ 的估计量可被给出为

$$\hat{g}_1(x, z_1) = \hat{c}_1 + \sum_{j=2}^{K_g+1} \hat{\beta}_j p_j(x, z_1)$$

$$\hat{g}_2(x, z_1) = \hat{c}_2 + \sum_{j=K_g+2}^{K} \hat{\beta}_j p_j(u) \quad (17.9)$$

其中 $\hat{c}_1 + \hat{c}_2 = \hat{\beta}_1$，且所有的系数除了常数项 $\hat{c}_1$ 和 $\hat{c}_2$ 外都是唯一被定义的。如果

对估计边际效应感兴趣的话，那么常数项就是不重要的。我们也可以估计常数项，但是因为修剪函数已经被使用，我们不能使用 $E(\varepsilon)=0$ 去识别它们，因为当 ε 被约束在修剪集内时，它将不再具有零均值。然而，我们可以引入其他的约束去帮助我们识别常数项。比如说，约束条件 $g_2(0)=0$ 可以帮助识别常数，因为 $E(Y|x,z_1,u=0)=g_1(x,z_1)+g_2(0)=g_1(x,z_1)$。这相当于 $\hat{c}_1=\hat{\beta}_1-\hat{c}_2$ 且 $\hat{c}_2=-\sum_{j=K_g+2}^{K}\hat{\beta}_j p_j(0)$。　525

下面的假定都被 Newey 等(1999)用于推导 $\hat{g}(w)$ 的渐近分布。

假定 17.1　$\Pi(z)$ 是 s_1 阶连续可微的，且 $g(x,z_1)$ 是 Lipschitz 且 s 阶连续可微的。

假定 17.2　对于幂级数，$((K^3+K^2L)(L/n)^{1/2}+L^{-s_1/d_1})\to 0$；对于样条，$((K^2+KL^{1/2})(L/n)^{1/2}+L^{-s_1/d_1})\to 0$。

Newey 等(1999)证明了下面的结论。

定理 17.1　在假定 17.1 和 17.2 以及某些可加正则性条件下，为样条设置 $q=1/2$ 且为幂级数设置 $q=1$，那么

(i) $\int\tau(w)[\hat{g}(w)-g(w)]^2\mathrm{d}F(w)=O_p((K/n)+K^{-2s/d}+(L/n)+L^{-2s_1/d_1})$

(ii) $\sup_{(x,z_1)\in W}|\hat{g}(w)-g(w)|=O_p(K^q[(K/n)^{1/2}+K^{-s/d}+(L/n)^{1/2}+L^{-s_1/d_1}])$

其中 $F(\cdot)$ 是 w 的累积分布函数。

收敛速度是这两项的和，且在第一阶段估计中依赖于序列数目 L，在第二阶段估计中依赖于 K。同时，(ii) 是由定理 15.1 而来的，其中对幂级数 $\zeta_0(K)=K$，且对样条 $\zeta_0(K)=K^{1/2}$。

为了计算 $\hat{g}(w)=p^K(w)'\hat{\beta}$ 的渐近方差，可以应用参数二阶段估计量的公式(如 Newey(1994a))。令 $\otimes$ 表示 Kronecker 乘积且定义 $\hat{Q}=\hat{P}'\hat{P}/n$，$\hat{\Sigma}_1Q_1^{-1}=\sum_{i=1}^{n}\hat{\tau}_i\hat{p}_i\hat{p}'_i[Y_i-\hat{g}(\hat{W}_i)]^2/n$，$\hat{Q}_1=I_{d-k_1}\otimes(R'R/n)$，$\hat{\Sigma}_1=\sum_{i=1}^{n}(\hat{u}_i\hat{u}_i')\otimes(r_ir_i')/n$，且 $\hat{H}=\sum_{i=1}^{n}\hat{\tau}_i[\partial\hat{g}(\hat{W}_i)/\partial u_i]'\otimes\hat{p}_ir'_i/n$，$\hat{g}(w)$ 的方差估计被给出为

$$\hat{V}(w)=p^K(w)'\hat{Q}^{-1}\left[\hat{\Sigma}+\hat{H}\hat{Q}_1^{-1}\hat{\Sigma}_1Q_1^{-1}\hat{H}'\right]\hat{Q}^{-1}p^K(w)\tag{17.10}$$

因为准确的偏误首项是很难得到的，Newey 等(1999)建议低度平滑这些数据，这导致偏误平方项比方差项有更小的阶，使得非参数序列估计值是恰当地处于中心的。特别是，他们假定当 $n\to\infty$ 时，　526

$$\sqrt{n}K^{-s/d}\to 0,\ \sqrt{n}L^{-s_1/d_1}\to 0\tag{17.11}$$

等式(17.11)保证了序列估计偏误平方项是小于方差项的。为了使得估计方差收敛于零,Newey 等(1999)假定

$$(K^9L + K^8L^2 + K^6L^3 + K^2L^6)/n \to 0, \text{幂级数}$$
$$(K^5L + K^4L^2 + K^3L^3 + KL^4)/n \to 0, \text{样条} \quad (17.12)$$

Newey 等(1999)证明了下面的渐近正态结果。

定理 17.2 在低度平滑条件(17.11)和(17.12)下,在某些关于 $g_1(\cdot)$,$\Pi(\cdot)$ 以及 $E(v_i^2 | X_i, Z_i)$ 的平滑性和矩条件下($v_i = \varepsilon_i$ 或 u_i),那么

$$\sqrt{n}\hat{V}(w)^{-1/2}(\hat{g}(w) - g(w)) \xrightarrow{d} N(0,1)$$

注意可以证明 $\sqrt{\max\{K,L\}}\hat{V}(w) = O_p(1)$。因此,式(17.2)意味着$((\hat{g}(w) - g(w))) = O_p(\sqrt{n}/\sqrt{\max\{K,L\}})$。

另外,可以使用自举法逼近 $\hat{g}(w)$ 的有限样本分布。事实上,Newey 等(1999)考虑了线性函数 $\hat{g}(\cdot)$ 的更一般的情形,它是 $\hat{\beta}$ 的线性函数(因为 $\hat{\beta}$ 关于 $\hat{g}(\cdot)$ 是线性的),比如,$\hat{\theta} = A\hat{\beta}$,其中 A 是一个 $m \times K$ 矩阵(m 是一个有限正常数)。$\hat{\theta}$ 的渐近分布为

$$\sqrt{n}\hat{V}_A^{-1/2}(\hat{\theta} - \theta) \xrightarrow{d} N(0, I_m) \quad (17.13)$$

其中 $\hat{V}_A = A\hat{Q}^{-1}[\hat{\Sigma} + \hat{H}\hat{Q}_1^{-1}\hat{\Sigma}_1\hat{Q}_1^{-1}\hat{H}']\hat{Q}^{-1}A'$,且 $\hat{V}_A$ 中所有的矩阵,除了 A,均在上面的式(17.10)中被定义。这个估计量和参数二阶段回归是类似的。它等于 $A\hat{Q}^{-1}\hat{\Sigma}\hat{Q}^{-1}A'$,一个异方差一致估计量,加上一个额外的非负确定性项,此项说明了 $\hat{u}$ 的出现。$\hat{\theta}$ 收敛于 θ 的速率依赖于 $\hat{V}_A$ 趋于无穷的速率。当 $A = p^K(w)'$时,式(17.13)推出了定理 17.2 。

527 17.3 Newey-Powell 的基于序列估计量

Newey 和 Powell(2003)考虑了一个更一般的联立方程模型:

$$Y_i = g(X_i, Z_{1i}) + \varepsilon_i, \quad E(\varepsilon_i | Z_i) = 0, \quad Z_i = (Z_{1i}, Z_{2i}) \quad (17.14)$$

其中 Y_i 为一个标量,且 X_i、Z_{1i} 和 Z_{2i} 的维度分别为 d_x、d_1 和 d_2。注意如果不进一步假定 $E(\varepsilon_i | Z_i, u_i) = E(\varepsilon_i | u_i)$,我们不能由式(17.14)得到类似式(17.15)的三角方程组。这会使得 $g_1(\cdot)$ 的一致估计更加困难。

式(17.14)的条件期望得出了下面的积分公式:

$$\Pi(z) \equiv E(Y_i | Z_i = z) = E[g(X_i, Z_{1i}) | Z_i = z] = \int g(x, z_1)F(\mathrm{d}x | z) \quad (17.15)$$

其中 F 表示给定 z 时 x 的条件累积分布函数。$g(\cdot)$ 的识别依赖于积分等式

(17.15)唯一解的存在。令 $\hat{\Pi}$ 和 $\hat{F}$ 表示 Π 和 F 的非参数估计量,我们需要基于下式得到 $g(\cdot)$ 的估计量:

$$\hat{\Pi}(z) = \int g(x, z_1) \hat{F}(\mathrm{d}x \mid z) \tag{17.16}$$

令 $\hat{g}(\cdot)$ 表示由式(17.16)得到的 $g(\cdot)$ 的一个解,在 $\hat{\Pi}$ 和 $\hat{F}$ 中 $\hat{g}(\cdot)$ 缺乏连续性说明约简式估计中的一个小的变动将导致 $\hat{g}(\cdot)$ 中大的变动。于是,与传统的估计量不同,$\hat{g}(\cdot)$ 的一致性不能由 $\hat{\Pi}$ 和 $\hat{F}$ 的一致性推出。这就是所谓的不适定逆(ill-posed inverse)问题。在关于积分等式的文献中,已经提出了许多方法处理非连续问题,且它们经常被称为"正则化方法"。Newey 和 Powell (2003)引入限制条件,$g(\cdot)$ 属于一个紧集,其中紧性是在 Sobolev 范数下定义的(见附录 A 的 A.31 关于其中范数的定义)。

令 $w=(x, z_1)$ 且令 $\{p_1(w), p_2(w), \cdots\}$ 为一个基础函数序列,Newey 和 Powell(2003)假定结构函数 $g(\cdot)$ 可以利用以下形式的某些序列基础函数的一个线性组合很好地逼近:

$$g_0(w) \cong g_J(w) = \sum_{j=1}^{J} \gamma_j p_j(x) \tag{17.17}$$

其中 γ_j 是对应于 $p_j(\cdot)$ 的系数,且 $p_j(\cdot)$ 是第 j 个被选择的(已知)基础函数(如多元或者三角函数),其性质为当 J 增大时,在均方误差意义上 $g_J(x)$ 可以任意好地逼近 $g(x)$。将式(17.17)代入式(17.14)可得 528

$$E(Y_i \mid Z_i) = E[g_0(W_i) \mid Z_i)] \cong \sum_{j=1}^{J} \gamma_j E[p_j(W_i) \mid Z_i] \tag{17.18}$$

公式(17.18)建议可以基于 Y_i 对 $E[p_j(W_i) \mid Z_i]$ 的最小二乘回归估计 $\{\gamma_j\}_{j=1}^{J}$。然而,这些条件期望函数都是未知的。Newey 和 Powell(2003)建议使用一个二阶段法,即在第一阶段利用一个非参数序列估计量估计$E[p_j(W_i) \mid Z_i]$,比如 $\hat{E}[p_j(W_i) \mid Z_i]$。然后,在第二阶段,用 Y_i 对 $\hat{E}[p_j(W_i) \mid Z_i]$ 的回归估计 γ_j。

然而,二阶段法可能遭遇先前所说的不适定逆问题。为了避免这个问题,Newey 和 Powell(2003)引入了限制条件,$g_0(w)$ 属于在 Sobolev 范数下的紧集。这个假定对于 $g_0(w)$ 的尾部行为施加了很强的限制。为了放松这一点,Newey 和 Powell(2003)假定 $g(\cdot)$ 函数有一个半参数形式$g(w)=a(w)'\beta+g_0(w)$,其中 $a(w)$ 和 β 分别是已知参数和未知参数的 $r\times 1$ 向量。这个设定在整个参数成分 $a(w)'\beta$ 允许无界的 w。如果对于一个给定的问题,w 是有界的,那么,我们可以丢掉参数部分并得到一个完全的非参数设定。对于半参数设定 $g(w)=$

$a(w)'\beta+g_0(w)$，式(17.16)变为

$$\hat{\Pi}(z) \cong \hat{E}(a(W)\mid z)'\beta+\sum_{j=1}^{J}\gamma_j\hat{E}(p_j(W)\mid z) \tag{17.19}$$

其中 $\hat{E}(a(W)|z)$ 和 $\hat{E}(p_j(W)|z)$ 分别是 $E(a(W)|z)$ 和 $E(p_j(W)|z)$ 的(一阶段)非参数序列估计量。在第二阶段，我们可以通过最小化下面的目标函数估计未知的 β 和 γ_j。

$$Q(\beta,\gamma)\stackrel{\text{def}}{=}\frac{1}{n}\sum_{i=1}^{n}\left[Y_i-\hat{E}(a(W_i)\mid Z_i)'\beta-\sum_{j=1}^{J}\gamma_j\hat{E}(p_j(W_i)\mid Z_i)\right]^2 \tag{17.20}$$

529 一个非参数二阶段最小二乘估计量可以在满足参数限制的条件下最小化式(17.20)，得到

$$\hat{g}(w)=a(w)'\hat{\beta}+\hat{g}_0(w),\ \hat{g}_1(w)=\sum_{j=1}^{J}\hat{\gamma}_jp_j(w)$$

其中

$$(\hat{\beta},\hat{\gamma})=\arg\min_{\beta,\gamma}Q(\beta,\gamma),\text{s. t.}\beta'\beta\leqslant B_\beta,\gamma'\Lambda_{J\gamma}\leqslant B_1$$

其中 B_β 和 B_1 都是某些正的常数，$\gamma'\Lambda_{J\gamma}$ 是线性逼近函数 g_J 的 Sobolev 范数的平方，且 Λ_J 是一个用包含基础函数 p_j 及其导数的积分来构造的已知矩阵。Newey 和 Powell(2003)建立了非参数估计量 $\hat{g}(w)$ 的均匀一致性，即几乎确定 $\sup_w|\hat{g}(w)-g(w)|=o(1)$，其中 sup 是在一个紧集内取值的，且概率密度函数 $f(w)$ 在紧集内存在一致上界和下界；参见 Newey 和 Powell(2003)关于用于证明 $\hat{g}(w)$ 对于 $g(w)$ 的均匀一致性的正则性条件的详细讨论。

17.4 Hall 和 Horowitz 的基于核的估计

接下来，非参数回归模型是与式(17.1)一样的，即

$$Y_i=g(X_i)+u_i \tag{17.21}$$

其中为了表达的简洁，我们首先考虑 X_i 是一个标量这种情况。因为 $E(u_i|X_i)\neq 0$，我们需要假定存在一个标量工具变量 Z_i，使得

$$E(u_i\mid Z_i)=0 \tag{17.22}$$

且 Z_i 和 X_i 是强相关的。然而，未知函数 $g(\cdot)$ 的估计是非平凡的，这正是我们接下来将要说明的。

令 f_x、f_z 和 f_{xz} 分别表示 X、Z 和 (X,Z) 的联合密度的边际密度。在积分平方函数空间定义如下的算子 T:

$$(T\psi)(w) = \int t(x,w)\psi(x)\mathrm{d}x \tag{17.23}$$

其中

530

$$t(x,w) = \int f_{xz}(x,z)f_{xz}(w,z)\mathrm{d}z \tag{17.24}$$

由式(17.23)和式(17.24),可以证明(见习题 17.1)

$$E\{E(Y \mid Z)f_{xz}(w,Z)\} = (Tg)(w) \tag{17.25}$$

如果我们假定 T 为非奇异的(即逆矩阵 T^{-1} 存在),那么,在式(17.25)两边同乘 T^{-1},可得

$$g(x) = E\{E(Y \mid Z)(T^{-1}f_{xz})(x,Z)\} \tag{17.26}$$

公式(17.26)表明,识别 $g(\cdot)$ 的关系是一个属于第一类的 Fredholm 方程:

$$Tg = \phi \tag{17.27}$$

其中 T 是一个线性算子,如果 0 是 T 的特征值的一个极限点,将会导致一个不适定逆问题。所以,T^{-1} 不是一个有界的连续算子。Hall 和 Horowitz(2003)建议使用一个脊型正则性(ridge-type regularization)去规避 T^{-1} 的非有界性,即用 $(T+a_n)^{-1}$ 取代 T^{-1},其中 a_n 是一个正脊参数,当 $n\to\infty$ 的其收敛于零。

我们现在考虑下面的多元情形,其模型为

$$Y_i = g(X_i, Z_{1i}) + u_i,\quad E(U_i \mid Z_{1i}, Z_{2i}) = 0 \tag{17.28}$$

其中 X 是内生的,Z 是外生的。Hall 和 Horowitz(2003)进一步假定 X 和 Z 具有相关联的紧支撑集,即 $X_i, Z_{2i} \in [0,1]^p$ 且 $Z_{1i} \in [0,1]^q$。

令 $k_h(u,t)$ 表示一个 r 阶的有界核函数,即如果 h 是小的且 t 不接近于 0 或者 1,那么 $k_h(u,t) = h^{-1}k(u/h)$。然而,如果 t 接近于 1,$k_h(u,t) = h^{-1}l(u/h)$,其中 $\int_0^\infty u^j l(u) = \delta_{0j}$ $j = 0,1$ (如果 $j = l, \delta_{jl} = 1$,否则为 0),且如果 t 接近于 0,$k_h(u,t) = l_h(-u/h)$。我们可以估计 (X,Z) 的联合密度 f_{xz} 为

$$\hat{f}_{xz}(x,z) = \frac{1}{n}\sum_{i=1}^{n} K_{p,h_x}(x - X_i, x)K_{q,h_z}(z_1 - Z_{1i}, z_1)K_{p,h_x}(z_2 - Z_{2i}, z_2)$$

其中 $K_{p,h_x}(u,t) = \prod_{s=1}^{p} h_{x,s}^{-1}k(u_s,t_s)$,$K_{q,h_z}(u,t) = \prod_{s=1}^{q} h_{z,s}^{-1}k(u_s,t_s)$,且其去一估计值为

531

$$\hat{f}_{xz,-i}(x,w,z) = \frac{1}{n-1}\sum_{i=1}^{n} K_{p,h_x}(x - X_i, x)K_{q,h_z}(z_1 - Z_{1i}, z_1)K_{p,h_x}(z_2 - Z_{2i}, z_2)$$

因此,可得

$$\hat{t}_{z_1}(x_1,x_2) = \int \hat{f}_{xz}(x_1,z)\hat{f}_{xz}(x_2,z)\mathrm{d}z$$

且

$$(\hat{T}_{z_1})(\psi)(x,z) = \int \hat{t}_{z_1}(\xi,x)\psi(\xi,z)\mathrm{d}\xi$$

其中 ψ 为一个映射$\mathbb{R}^{2p+q}$到实线(real line)的函数。那么,$g(x,z_1)$的估计量为

$$\hat{g}(x,z_1) = \frac{1}{n}\sum_{i=1}^{n}\hat{T}_{z_2}^{+}\hat{f}_{xz,-i}(x,z)Y_i K_{q,h_z}(z_2 - Z_{2i},z_2)$$

因为 $g(\cdot)$的一致估计依赖于 T 的可逆性,$\hat{g}(\cdot)$的收敛速度依赖于 T 的特征值收敛于零的速率。令 $\varphi_1,\varphi_2,\cdots$表示积分平方函数在[0,1]上的正交基,那么,$t(\cdot)$,$f_{XW}$和 $g(\cdot)$有下面的扩展:

$$t(x,z) = \sum_{j=1}^{\infty}\lambda_j\phi_j(x)\phi_j(z)$$

$$t(x,z) = \sum_{j=1}^{\infty}\sum_{l=1}^{\infty}d_{jl}\phi_j(x)\phi_l(z)$$

$$g(x) = \sum_{j=1}^{\infty}b_j\phi_j(x)$$

其中 d_{jl}和 b_j 分别是f_X 和 $g(\cdot)$的广义傅里叶系数。

令 α 和 β 为满足 $\alpha>1$ 的常数。其中 $\beta>1/2$ 且 $\beta-1/2\leqslant\alpha<2\beta$,其中 $|b_j|\leqslant Cj^{-\beta}$,$j^{-\alpha}\leqslant C\lambda_j$ 且 $\sum_{l\geqslant j}|d_{jl}|\leqslant Cj^{-\alpha/2}$,对所有 $j>1$ 成立,Hall 和 Horowitz (2003)证明了下面的一致收敛速率结果:

532

$$\sup_{G\in\mathcal{M}}\int_{[0,1]^q}E_G\{\hat{g}(t)-g(t)\}^2\mathrm{d}t = O(n^{-(2\beta-1)/(2\beta+\alpha)})$$

令 $\tau=2r/(2r+q)$,Hall 和 Horowitz(2003)进一步证明在下面的最小化意义上的最优收敛速率为 $n^{-\tau(2\beta-1)/(2\beta+\alpha)}$。

$$\liminf_{n\to\infty}n^{\tau(2\beta-1)/(2\beta+\alpha)}\inf_{\hat{g}}\sup_{G\in\mathcal{C}}\int_{[0,1]^q}E_G\{\hat{g}(x,z_1)-g(x,z_1)\}^2\mathrm{d}x > 0$$

在稍微更强的条件下,包括要求 α 以比渐近最优速率更快的速度收敛于零,同时将平滑参数 h 限制在一个稍微更窄的范围内,Horowitz(2005)建立了 $\hat{g}(x,z_1)$的逐点渐近正态性结果。

17.5 Darolles,Florens 和 Renault 的估计量

Darolles,Florens 和 Renault(2002)考虑了相同的非参数回归模型,即

$$Y = g(X)+u,\ E(u\mid Z) = 0 \tag{17.29}$$

其中 Z 为工具变量。因此,$g(\cdot)$被定义为函数方程的解

$$E(Y - g(X) \mid Z) = 0 \tag{17.30}$$

为了估计未知函数 $g(\cdot)$，Darolles 等（2002）开始注意 $g(\cdot)$ 是下面的方程的解：

$$g(\cdot) = \arg\min_{\phi(\cdot)} E[\| E(Y|z) - \int \phi(x) F(\mathrm{d}x|z) \|^2]$$

接下来，用 $T_F(g) = E(g(x)|z)$ 和 $T_F^*(\psi) = E(\psi(x)|z)$ 定义两个线性算子 $T_F(g)$ 和 $T_F^*(\psi)$。使用这个记号，$g(\cdot)$ 对应于任何函数方程的解

$$A(g,F) = T_F(g) - r_F = 0 \tag{17.31}$$

其中 $r_F(z) = E(Y|z)$。令 $f(y,x,z)$ 为 (Y,X,Z) 的概率密度函数，那么式（17.31）是一个积分的 Fredholm 类型 I 公式：

533

$$\int g(x) \frac{f(\cdot,x,z)}{f(\cdot,\cdot,z)} \mathrm{d}x - \tau_F(z) = 0 \tag{17.32}$$

其中 $r_F(z) = \int y \frac{f(y,\cdot,z)}{f(\cdot,\cdot,z)} \mathrm{d}y$。

为了识别 $g(\cdot)$，Darolles 等（2002）引入条件几乎确定 $E(g(X)|Z) = 0$，这意味着几乎确定 $g(X) = 0$。在这个条件下，$g(\cdot)$ 是可以识别的。令 $\phi_j, j \geqslant 1$ 为 $L_F^2(z)$ 的正交序列，即 $\int \phi_j(w)\phi_l(w)\mathrm{d}F(w) = \delta_{jl}$。对 $j \geqslant 1$，$T_F(\phi_j) = \lambda_j \psi_j$，$T_F^*(\psi_j) = \lambda_j \phi_j$，$\phi_1 = 1$，$\psi_1 = 1$ 且 $\int \phi_j(x)\phi_l(x)\mathrm{d}F(z,x) = \lambda_j \delta_{jl}$，可以证明（见习题 17.2）

$$T_F[g(X)](z) = E[g(X) \mid z] \sum_{j=1}^{\infty} \lambda_j < g(x), \phi_j(x) > \psi_j(z) \tag{17.33}$$

那么，

$$g(x) = \sum_{j=1}^{\infty} \frac{1}{\lambda_j} < r_F, \psi_j > \phi_j(x)$$

其不适定逆问题对应于当 $j \to \infty$ 时 $\lambda_j \to 0$。小的估计误差将会导致估计 $g(\cdot)$ 函数的极大改变。Darolles 等（2002）建议使用一个“切断估计量”（cutoff estimator），其只使用 ϕ_j 和 ψ_j，有 $\lambda_j^2 \geqslant \alpha_n$，其中 $\alpha_n > 0$ 是一个正数序列，当 $j \to \infty$ 时，$\alpha_j \to 0$，因此

$$\hat{g}_{\alpha_n}(x) = (\alpha_n I + T_{\hat{F}_n}^* T_{\hat{F}_n})^{-1} r_{\hat{F}_n}^*$$

其中 $r_{\hat{F}_n}^* = T_{\hat{F}_n}^* r_{\hat{F}_n}$，且 $\hat{F}_n$ 是核平滑经验累积分布函数。Darolles 等提供了 $\hat{g}_{\alpha_n}(w)$ 到 $g(w)$ 的收敛速率和 $\hat{g}_{\alpha_n}(w)$ 的一个渐近正态结果。

这种方法已经被 Blundell，Chen 和 Kristensen（2003）使用，他们考虑了在出

现外生支出的情形下半参数恩格尔曲线的估计。

17.6 习 题

习题 17.1 由式(17.23)和(17.24)推导式(17.25)。

习题 17.2 推导式(17.33)。

第18章　弱依赖数据

迄今为止，我们把注意力限制在独立数据上。在这一章，我们证明所有的早期研究结果都可以扩展到平稳弱依赖数据体系。接下来将要提到的"弱依赖"的意思是指当 τ 趋于无穷时，Z_t 和 $Z_{t+\tau}$ 之间的依赖性为0。此外，当 $\tau\to\infty$ 时，我们也需要强加一定的收敛速率。于是，弱依赖数据排除了非平稳进程的可能性（比如单位根），或者过程拥有所谓的远期依赖。在这一章，我们主要关注严格平稳弱依赖数据。在18.9节我们简要地讨论关于非平稳数据回归模型的非参数估计。为了描述弱依赖过程，我们使用各种所谓的混合过程，我们在下面给出定义。 535

令 $\mathcal{M}_t^{t+\tau}$ 表示由 $\{Z_s\}_{s=t}^{t+\tau}$ 产生的 σ 代数，令 $\mathbb{N}$ 表示整数集，且定义

$$\phi_\tau = \phi(\tau) = \sup_{t\in\mathbb{N}} \sup_{A\in\mathcal{M}_{t+\tau}^{\infty}, B\in\mathcal{M}_{-\infty}^{t}} |P(A|B) - P(A)|$$

$$\alpha_\tau = \alpha(\tau) \sup_{t\in\mathbb{N}} \sup_{A\in\mathcal{M}_{t+\tau}^{\infty}, B\in\mathcal{M}_{-\infty}^{t}} |P(A\cap B) - P(A)P(B)|$$

$$\beta_\tau = \beta(\tau) \sup_{t\in\mathbb{N}} \sup_{A\in\mathcal{M}_{t+\tau}^{\infty}, B\in\mathcal{M}_{-\infty}^{t}} |E(A|B) - E(A)|$$

$$\rho_\tau = \rho(\tau) \sup_{t\in\mathbb{N}} \sup_{A\in\mathcal{M}_{t+\tau}^{\infty}, B\in\mathcal{M}_{-\infty}^{t}} \left|\frac{\text{cov}(A,B)}{\sqrt{\text{var}(A)\text{var}(B)}}\right| \tag{18.1}$$

序列 $\{Z_t\}_{t=-\infty}^{\infty}$ 可以说是 ϕ 混合，或 α 混合，或 β 混合，或 ρ 混合，如果当 $\tau\to\infty$ 时，ϕ_τ、α_τ、β_τ 和 ρ_τ 分别趋于零。注意 α-混合也可以称为"强混合"。

注意对于独立的数据，对 $\tau\geqslant 1$，ϕ_τ、α_τ、β_τ 和 ρ_τ 都是0。混合过程可以被看作是渐近独立过程。 536

可以证明

$$\alpha_\tau \leqslant \beta_\tau \leqslant \phi_\tau, \alpha_\tau \leqslant \rho_\tau \leqslant \phi_\tau \tag{18.2}$$

然而，一般情形下我们不能对 β_τ 和 ρ_τ 进行排序。

参见 Carrasco 和 Chen(2002),以及 Chen,Hansen 和 Scheinkman(2005)关于不同的混合过程性质的进一步讨论。

Ibragimov 和 Linnik(1971)证明一个高斯自回归移动平均 ARMA(p,q)过程($p,q \in \mathbb{N}_+$)就是这样的过程,当 $\tau \to \infty$ 时,$\alpha_\tau \to 0$,但是 ϕ_τ 不是。此外,$\alpha_\tau \to 0$ 以指数的速度收敛。[①] 也许混合过程最方便的性质在于混合过程的可测函数本身是混合性质的,给定函数仅仅依赖于混合过程的一个有限的滞后期数。然而,当 Y_t 依赖于潜在过程的整个历史时,Y_t 的可测函数可能不是一个混合过程,使 Y_t 本身是混合的。在这种情况下,混合过程的概念被 McLeish(1975)扩展为"混合鞅"(mixingale)过程。关于混合鞅的细节处理和其他的弱依赖过程,我们请读者参考 Gallant 和 White(1988)以及 Andrews(1988)。

在一篇核估计的重要论文中,Robinson(1983)建立了对于非参数密度,回归和具有强混合(α 混合)数据的条件密度估计量的一致性和渐近正态性。因为其他的三个混合过程意味着 α 混合,α 混合过程因此是在上述提及的弱依赖过程中依赖最强的过程。

在这一章我们把注意力集中于平稳的 ρ 混合或 β 混合过程,因为从数学上来说,ρ 混合和 β 混合是最容易处理的。比如说,一个标准的 AR(1)模型的形式为 $Y_t = \alpha Y_{t-1} + u_t$,其中 u_t 是一个具有有限方差 σ^2 的白噪声误差项,是 ρ 混合,给定 $|\alpha| < 1$,因为当 $\tau \to \infty$ 时,$\rho_\tau = \mathrm{cov}(Y_{t+\tau}, Y_t) / \sqrt{\mathrm{var}(Y_t)\mathrm{var}(Y_{t+\tau})} = \alpha^\tau \to 0$。Carrasco 和 Chen(2002)证明许多已知的非线性时间序列过程都是 β 混合(或 ρ 混合)。

我们将提供密度和回归函数的核估计量的渐近正态性和逐点一致性的详细证明。我们还要讨论许多半参数模型的估计以及对于弱依赖数据的一致模型设定检验。

537

这一章的主要目的是说明在前面的章节中得到的关于独立性假定的结论,如收敛速率和渐近正态性,当独立性假定被放松为允许弱依赖时,都不会改变。

18.1 关于依赖数据的密度估计

为了表达的简洁,我们只考虑 $x \in \mathbb{R}^q$ 是向量一值连续变量这种情形。将这一结论扩展到混合离散和连续变量的情形是相当直接的,正如我们在第 4 章中所做的那样。假定我们有平稳性和 ρ 混合观测值 $X_1, X_2, \cdots, X_n$。正如在第 1 章那样,对于一个给定的 $x \in \mathbb{R}^q$ 的固定值,我们估计 X_t 的概率密度函数 $f(x)$ 为

$$\hat{f}(x) = \frac{1}{n} \sum_{t=1}^{n} K_h(X_t - x)$$

---

① 对于非高斯过程,Andrews(1984)证明了一个简单的 AR(1)过程可以既不是 ϕ 混合也不是 α 混合。

其中，$K_h(X_t - x) = \prod_{s=1}^{q} h_s^{-1} k((X_{ts} - x_s)/h_s)$ 是乘积核。

接下来的定理证明具有弱依赖数据的 $\hat{f}(x)$ 的均方误差收敛速率与我们在第 1 章中讨论的独立数据情形是一样的。

定理 18.1　假定一个 ρ 混合过程满足 $\rho(\tau) = O(\tau^{-(1+\varepsilon)})$，某些（小的）$\varepsilon > 0$（这意味着 $\sum_{\tau=1}^{\infty} \rho(\tau) < \infty$），在定理 1.3 中给出的关于 $f(\cdot)$ 和 $k(\cdot)$ 相同的平滑和矩条件成立，且当 $n \to \infty$ 时，$h_s \to 0\ (s = 1, \cdots, q)$，$nh_1 \cdots h_q \to \infty$，那么 $\left(\| h \|^2 = \sum_{s=1}^{q} h_s^2 \right)$

$$E\{[\hat{f}(x) - f(x)]^2\} = O(\| h \|^4 + (nh_1 \cdots h_q)^{-1})$$

证明：令 $K_{h,tx} = K_h(X_t - x) = \prod_{s=1}^{q} h_s^{-1} k((X_{ts} - x_s)/h_s)$，那么，由平稳性假定，可得

538

$$\begin{aligned} \text{bias}(\hat{f}(x)) &= E[\hat{f}(x)] - f(x) = E(K_{h,1x}) - f(x) \\ &= \int K_h(x_1 - x) f(x_1) \mathrm{d}x_1 - f(x) \\ &= \int K(v) f(x + hv) \mathrm{d}v - f(x) \\ &= (1/2)\kappa_2 \sum_{s=1}^{q} h_s^2 f_{ss}(x) + O\left(\sum_{s=1}^{q} h_s^3 \right) \end{aligned} \tag{18.3}$$

注意上面的偏误计算和独立同分布数据的情形是完全相同的，因为平稳性就意味着相同的分布。

接下来，我们计算方差项，给出为：

$$\begin{aligned} \text{var}(\hat{f}(x)) &= n^{-2} \text{var}\left(\sum_{t=1}^{n} K_{h,tx} \right) \\ &= n^{-2} \left\{ \sum_{t=1}^{n} \text{var}(K_{h,tx}) + 2 \sum_{t=1}^{n} \sum_{s>t}^{n} \text{cov}(K_{h,tx}, K_{h,sx}) \right\} \end{aligned} \tag{18.4}$$

下面我们分别计算 $\text{var}(K_{h,tx})$ 和 $\text{cov}(K_{h,tx}, K_{h,\cdot,sx})$。$\text{var}(K_{h,tx})$ 的计算和独立同分布数据的情形下是一样的，在平稳性条件下可得（用 $H_q = h_1 \cdots h_q$ 表示）

$$\begin{aligned} \text{var}(K_{h,tx}) &= E[K_{h,tx}^2] - [E(K_{h,tx})]^2 \\ &= \int K_h^2(x_t - x) f(x_t) \mathrm{d}x_t - [\int K_h(x_t - x) f(x_t) \mathrm{d}x_t]^2 \\ &= H_q^{-1} \int K^2(v) f(x + hv) \mathrm{d}v - [\int K(v) f(x + hv) \mathrm{d}v]^2 \\ &= H_q^{-1} f(x) \int K^2(v) \mathrm{d}v + O\left(H_q^{-1} \sum_{s=1}^{q} h_s^2 + 1 \right) \end{aligned} \tag{18.5}$$

根据平稳性性质,式(18.5),并假定ρ混合,可得

539

$$
\begin{aligned}
|\operatorname{cov}(K_{h,tx},K_{h,sx})| &\leqslant \rho(s-t)\operatorname{var}(K_{h,tx}) \\
&= H_q^{-1}\rho(s-t)f(x)\int K^2(v)\,\mathrm{d}v + O\left(H_q^{-1}\sum_{s=1}^{q}h_s^2+1\right) \quad (18.6)
\end{aligned}
$$

使用 $\sum_{t=1}^{n-1}\sum_{s>t}^{n}\rho(s-t) = \sum_{t=1}^{n-1}\sum_{j=1}^{n-t}\rho(j) \leqslant n\sum_{j=1}^{\infty}\rho(j)$(其中 $j=s-t$),可得

$$
\begin{aligned}
\operatorname{var}(\hat{f}(x)) &= n^{-2}\left\{\sum_{t}\operatorname{var}(K_{h,tx}) + 2\sum_{t}\sum_{s>t}\operatorname{cov}(K_{h,tx},K_{h,sx})\right\} \\
&\leqslant n^{-2}\left\{n\operatorname{var}(K_{h,tx}) + 2n\operatorname{var}(K_{h,tx})\sum_{j=1}^{\infty}\rho(j)\right\} \\
&= O((nh_1\cdots h_q)^{-1}) \quad (18.7)
\end{aligned}
$$

其中,最后一个等式利用了式(18.5)的 $\operatorname{var}(K_{h,tx}) = O((h_1\cdots h_q)^{-1})$,以及 $\sum_{j=1}^{\infty}\rho(j) < \infty$ 的事实。

综合式(18.3)和式(18.7),可得

$$
\operatorname{MSE}(\hat{f}(x)) = O\left(\left(\sum_{s=1}^{q}h^2\right)^2 + (nh_1\cdots h_q)^{-1}\right) \quad (18.8)
$$

我们已经证明,对于弱依赖过程,核密度估计量$\hat{f}(x)$与独立同分布情形下的均方误差具有同样的阶数。事实上,我们将要证明$\hat{f}(x)$的渐近分布也是与独立同分布情形下相同的,特别是,这意味着偏误首项和方差首项与独立同分布背景下是一样的。

在18.10.1节我们证明

$$
\operatorname{var}((nh_1\cdots h_q)^{1/2}\hat{f}(x)) = \kappa^q f(x) + o(1) \quad (18.9)
$$

公式(18.9)表明对于ρ混合过程来说,它的混合系数满足 $\sum_{\tau=1}^{\infty}\rho(\tau) < \infty$,$\hat{f}(x)$ 的渐近方差首项是与独立同分布数据情形下相同的。

540

给定$\hat{f}(x)$的偏误首项和方差首项与独立同分布数据情形下是相同的,很自然地,我们期望逐点渐近正态性结论对于混合数据也是成立的。接下来的定理给出了这一结论。

定理18.2 在与定理2.2给出的关于$f(\cdot)$和$k(\cdot)$相同的条件下,但是把独立同分布假定用具有$\rho(\tau)=O(\tau^{-(1+\varepsilon)})$的$\rho$混合或具有$\alpha(\tau)=O(\tau^{-(1+\varepsilon)})$的$\alpha$混合替代,并且也假定当$n\to\infty$时$h_s\to 0(s=1,\cdots,q)$,$nh_1\cdots h_q\to 0$,且 $nh_1\cdots h_q\sum_{s=1}^{q}h_s^6\to\infty$,然后可得

$$\sqrt{nh_1\cdots h_q}\left(\hat{f}(x)-f(x)-\frac{1}{\kappa_2}\sum_{s=1}^{q}h_s^2 f_{ss}(x)\right)\xrightarrow{d}N(0,\kappa^q f_J(x))$$

在 ρ 混合下定理 18.2 的证明在 18.10 节给出。在 α 混合下的证明可参见 Robinson(1983)和 Masry(1996b)。

定理 18.2 说明对于弱依赖过程 $\hat{f}(x)$ 的(一阶)渐近分布是与独立同分布数据情形下相同的。

我们也可以考虑在 m 个不同点求值以估计一个联合概率密度函数向量。令 $\boldsymbol{x}=(x_{1+\tau_1},\cdots,x_{1+\tau_m})$,其中 $0\leqslant\tau_1<\tau_2<\cdots<\tau_m$ 是一些正整数。定义一个 $m\times1$ 向量函数 $f_J(\boldsymbol{x})$ 为

$$f_J(\boldsymbol{x})\overset{\text{def}}{=}(f(x_{1+\tau_1}),f(x_{1+\tau_2}),\cdots,f(x_{1+\tau_m}))'$$

定义 $f_J(\boldsymbol{x})$ 的核估计为

$$\hat{f}_J(x)\overset{\text{def}}{=}(\hat{f}(x_{1+\tau_1}),\hat{f}(x_{1+\tau_2}),\cdots,\hat{f}(x_{1+\tau_m}))'$$

那么可以证明 $\hat{f}_J(\boldsymbol{x})$ 收敛于 $f_J(\boldsymbol{x})$。的确,我们可以证明

$$\sqrt{nH^q}\left(\hat{f}_J(\boldsymbol{x})-f_J(\boldsymbol{x})-\frac{1}{\kappa_2}\sum_{s=1}^{q}h_s^2 f_{J,ss}(\boldsymbol{x})\right)\xrightarrow{d}N(0,\kappa^q f(x))\qquad(18.10)$$

其中 $f_{J,ss}(\boldsymbol{x})$ 是一个 $m\times1$ 向量,其第 j 个元素是

$$f_{ss}(x_{1+\tau_j})=[\partial^2 f(x)/\partial x_s^2]\,|_{x=x_{1+\tau_j}}$$

我们观察到 $\hat{f}(x_{1+\tau_j})$ 对于不同的 j 是渐近独立的。Robinson(1983)提供了一个式(18.10)的关于 α 混合数据的证明。　541

18.1.1　几乎确定一致收敛速率

正如独立同分布数据的情形,对于弱依赖数据的几乎确定一致收敛速率没有改变。

定理 18.3　在与定理 1.4 给出的关于 $f(\cdot)$ 和 $k(\cdot)$ 的相似条件下,且假定数据是 ρ 混合的,$\rho(\tau)=O(\tau^{-(1+\varepsilon)})$,或者 α 混合的,$\alpha(\tau)=O(\tau^{-(1+\varepsilon)})$,那么

$$\sup_{x\in\mathcal{D}}|\hat{f}(x)-f(x)|=O\left(|h|^2+\left(\frac{\ln(n)}{nh_1\cdots h_q}\right)^{1/2}\right)\text{ a. s.}$$

证明:见 Masry(1996b)。□

定理 18.3 告诉我们 $\hat{f}(x)$ 的几乎确定一致速率是与独立同分布数据情形下相同的(见定理 1.4)。

18.2 依赖数据的回归模型

18.2.1 鞅差分误差情形

考虑下列形式的非参数回归模型：

$$Y_t = g(X_t) + u_t \tag{18.11}$$

其中，正如第2章所述，$g(\cdot)$是一个未知的平滑函数，且我们假定u_t是一个鞅差分过程（因此u_t是序列无关的）。假定$\{Y_t, X_t'\}_{t=1}^n$是一个严格平稳的β混合过程，其混合系数是β_τ。我们用局部常数核方法估计$g(x)$，因此

$$\hat{g}(x) = n^{-1}\frac{\sum_{t=1}^{n} Y_t K_h(x - X_t)}{\hat{f}(x)} \tag{18.12}$$

我们将假定误差项u_t是一个鞅差分过程，即假定$E(u_t | X_t, \mathcal{M}_{-\infty}^{t-1}) = 0$（对所有$\tau \geqslant 1$，$\mathcal{M}_{-\infty}^{t-1}$包含$(Y_{t-\tau}, X_{t-\tau})$）。鞅差分假定意味着误差$u_t$是序列无关的，但是它允许误差为未知的条件异方差形式。同时，当X_t包含Y_t的滞后值时，如$Y_{t-1}, \cdots, Y_{t-d}$，那么$E(u_t | X_t) = 0$意味着$E(u_t | Y_{t-1}, \cdots, Y_{t-d}) = 0$，这说明$u_t$可能是序列无关的。因此，关于$u_t$的鞅差分条件对非参数时间序列回归模型是一个合理的假定。

542

为了推导$\hat{g}(x)$的渐近分布，就像我们在处理独立同分布数据情形的那样，可写出

$$\hat{g}(x) - g(x) = [\hat{g}(x) - g(x)]\hat{f}(x)/\hat{f}(x) \equiv \hat{m}(x)/\hat{f}(x)$$

其中

$$\hat{m}(x) = [\hat{g}(x) - g(x)]\hat{f}(x) = \hat{m}_1(x) + \hat{m}_2(x)$$

$$\hat{m}_1(x) = n^{-1}\sum_t [g(x_t) - g(x)]K_{h,tx}$$

$$\hat{m}_2(x) = n^{-1}\sum_t u_t K_{h,tx}$$

用与独立同分布数据模型的情形下完全一样的证明，我们可以证明（$|h|^j = \sum_{s=1}^{q} h_s^j$）

$$E[\hat{m}_1(x)] = \sum_{s=1}^{q} h_s^2 B_s(x) + O(\|h\|^3) \tag{18.13}$$

其中$B_s(x) = (1/2)\kappa_2[f(x)g_{ss}(x) + 2f_s(x)g_s(x)]$。

使用与求 $\mathrm{var}(\hat{f}(x))$ 时一样的计算，我们可以证明（见习题 18.2）

$$\mathrm{var}[\hat{m}_1(x)] = O(|h|^2(nH_q)^{-1}) = o((nH_q)^{-1}) \tag{18.14}$$

公式（18.13）和（18.13）意味着

$$\hat{m}_1(x) = \sum_{s=1}^{q} h_s^2 B_s(x) + o\left(|h|^2 + (nH_q)^{-1/2}\right) \tag{18.15}$$

假定 u_t 遵循一个鞅差分过程，对于 $t \neq s$，我们有 $E(u_t u_s | X_t, X_s) = 0$，因此，由与独立同分布数据情形下完全相同的证明可以得到（见第 2 章）

$$\begin{aligned}
E[(\hat{m}_2(x))^2] &= n^{-2}\sum_t \sum_s E[u_t u_s K_h(x - X_t)K_h(x - X_s)] \\
&= n^{-2}\sum_t E[u_t^2 K_h^2(x - X_t)]\ (\text{因为 } E(u_t u_s | X_t, X_s) = 0, \text{如果 } t \neq s) \\
&= n^{-2} E[\sigma^2(X_t)K_h^2(x - X_t)] \\
&= (nh_1 \cdots h_q)^{-1}\{\kappa^q \sigma^2(x) f(x) + o_p(1)\}
\end{aligned} \tag{18.16}$$

543

公式（18.16）意味着（$H_q = h_1 \cdots h_q$）

$$\hat{m}_2(x) = O_p((nH_q)^{-1/2}) \tag{18.17}$$

由定理 18.1 和式（18.17）推出（假定 $f(x) > 0$）

$$\begin{aligned}
\hat{g}(x) - g(x) &= \frac{\hat{m}(x)}{\hat{f}(x)} \\
&= \frac{O_p(|h|^2 + (nH_q)^{-1/2})}{f(x) + o_p(1)} \\
&= O_p(|h|^2 + (nH_q)^{-1/2})
\end{aligned} \tag{18.18}$$

进一步我们可以建立 $\hat{g}(x)$ 的渐近分布。

定理 18.4　假定当 $n \to \infty$ 时，$h \to 0$，$nh^q \to \infty$，$nh^{q+8} \to 0$，可得

$$\sqrt{nH_q}\left(\hat{g}(x) - g(x) - \sum_{s=1}^{q} h_s^2 B_s(x)\right) \xrightarrow{d} N(0, \kappa^q \sigma^2(x)/f(x))$$

我们可以使用和定理 18.2 的证明同样的“小方块”（small block）和“大方块”方法去证明定理 18.4（详见下面的 18.10 节）；因此，我们省略此处的证明过程。

定理 18.4 说明 $\hat{g}(x)$ 的（一阶）渐近分布与独立同分布数据情形下是一样的。

由式（18.16）的推导，我们可以看出鞅差分误差假定使得渐近方差的计算与独立数据情形下一样简单。定理 18.4 在没有鞅差分假定的前提下仍然成立，给定误差 u_t 是一个混合过程，且混合系数满足一个衰变率，正如在定理 18.2 的证明中所使用的那样。

正如独立同分布数据的情形，我们也可以用最小二乘交错鉴定方法选择平滑参数，即通过选择 $h_1,\cdots,h_q$ 最小化下式：

$$\mathrm{CV}(h) = n^{-1}\sum_{t=1}^{n}[Y_t - \hat{g}_{-t}(X_t)]^2 M(X_t)$$

544

其中 $\hat{g}_{-t}(X_t) = \sum_{s\neq t}^{n} Y_s K_h(X_s - X_t) / \sum_{s\neq t}^{n} K_h(X_s - X_t)$ 是 $g(X_t)$ 的去一核估计量，并且 $M(\cdot)$ 是一个紧支撑的非负加权函数。在数据是混合过程，其混合系数以某个速率衰变为零的假定下，可以证明交错鉴定法选择平滑参数与独立数据情形有同样的渐近表现（比如，见第 2 章）。这并不是令人惊讶的，因为我们已经证明，不管数据是独立的还是弱依赖的，均方误差估计都有同样的首项展开。

18.2.2 自相关误差情形

Xiao，Linton，Carroll 和 Mammen(2003) 考虑了以下形式的回归模型：

$$Y_t = g(X_t) + u_t,\ t = 1,\cdots,n \tag{18.19}$$

其中 $X_t \in \mathbb{R}^q$ 是严格外生的，即 $E(u_t | X_1,\cdots,X_n) = 0$ 对于所有的 t 成立，Xiao等允许 u_t 为序列相关的，我们写为

$$u_t = \sum_{j=0}^{\infty} c_j \varepsilon_{t-j} \tag{18.20}$$

其中 ε_j 服从具有零均值和有限方差 σ_ε^2 的独立同分布，且 c_j 是未知系数。注意 Xiao 等假定 X_t 和 u_t 是正交的，即 $E(u_t|X_t) = 0$。因为 u_t 是序列相关的，这意味着 X_t 不能包含 Y_t 的滞后值 。令 L 表示滞后算子，且 $c(L) = \sum_{j=0}^{\infty} c_j L^j$，$c(L)$ 的逆为

$$c(L)^{-1} = a(L) = a_0 - a_1 L - \cdots - a_j L^j - \cdots = a_0 - \sum_{j=1}^{\infty} a_j L^j$$

将 $a(L)$ 应用到式(18.19)推出

$$a(L)Y_t = a(L)g(X_t) + \varepsilon_t \tag{18.21}$$

式(18.21)中的误差项是不相关的。定义

$$\underline{Y}_t = Y_t - \sum_{j=1}^{\infty} a_j (Y_{t-j} - g(X_{t-j}))$$

那么式(18.21)可被重写为

$$\underline{Y}_t = g(x_t) + \varepsilon_t \tag{18.22}$$

545

公式(18.22)是一个有效的回归方程，因为误差 ε_t 是序列不相关的。Xiao 等(2003)首次考虑了基于 $\underline{Y}_t$ 是已知的假定下 $g(\cdot)$ 的一个不可行估计量。Xiao 等考虑了一般的局部多项式方法。这里，为了叙述简便，我们只考虑局部线性方法。令 $\bar{g}(x)$ 表示 $g(x)$ 基于式(18.22)的局部线性估计量，那么在某些附加的假定下，比如数据是 α 混合的，其混合系数满足一个特定的衰变率，Xiao 等

证明

$$\sqrt{nh_1\cdots h_q}\left(\bar{g}(x)-g(x)-\frac{1}{\kappa_2}\sum_{s=1}^{q}g_{ss}(x)\right)\xrightarrow{d}N(0,\sigma_\varepsilon^2\kappa^q/f(x))$$

(18.23)

其中，$f(\cdot)$是 X_t 的概率密度函数。注意式(18.23)和我们在第 2 章中讨论的独立数据情形是一样的。同样注意 $\bar{g}(x)$是比基于式(18.19)的估计量更有效的，因为 $\sigma_\varepsilon^2<\sigma_u^2=\sigma_\varepsilon^2\sum_{j=0}^{\infty}c_j^2$。Xiao 等进一步考虑了一个可行的估计过程，其过程描述如下：(i) 基于式(18.19)估计 $g(\cdot)$(比如 $\hat{g}(X_t)$)，(ii) 得到 $\hat{u}_t=Y_t-\hat{g}(X_t)$，然后估计 $\hat{u}_t$ 的一个 τ 阶自回归，即 $\hat{u}_t=\hat{a}_1\hat{u}_{t-1}+\cdots+\hat{a}_\tau\hat{u}_{t-\tau}+\text{residual}$，(iii) 使用 $\hat{\underline{Y}}_t=Y_t-\sum_{t=1}^{\tau}\hat{a}_j(Y_{t-j}-\hat{g}(X_{t-j}))$ 构造到 $\underline{Y}_t$ 的一个近似，最后(iv) 基于 $\hat{\underline{Y}}_t=g(X_t)+\text{error}$ 再次估计 $g(\cdot)$。这个过程可以迭代以得到更好的有限一样本表现。令 $\tilde{g}(x)$ 表示 $g(\cdot)$ 的最终的可行估计量。Xiao 等证明 $\tilde{g}(x)$ 与式(18.23)中描述的 $\bar{g}(x)$ 有同样的渐近分布。

我们可以简单地允许误差项遵循一个有限阶非参数 ARCH 过程，比如 $\text{var}(u_t|Y_{t-1},\cdots,Y_{t-p})=m(Y_{t-1},\cdots,Y_{t-p})$，其中 $m(\cdot)$的函数形式是未设定的。可以用 $\hat{u}_t^2$ 对于$(Y_{t-1},\cdots,Y_{t-p})$的非参数回归来一致估计 $m(\cdot)$，其中 $\hat{u}_t$ 是估计残差；关于非参数 ARCH(p)模型的理论和应用参见 Pagan 和 Schwert(1990)以及 Härdle 和 Tsybakov(1997)。估计 GARCH 或 ARCH(∞)模型引入了一个更加困难的问题。最近，Linton 和 Mammen(2005)提出了一个新颖的方法来一致估计形式为 $\text{var}(Y_t|I_{t-1})=\sum_{j=1}^{\infty}\psi_j(\theta)m(Y_{t-j})$的半参数 ARCH($\infty$)模型，其中 θ 是一个有限维度参数且 $m(\cdot)$的函数形式是未设定的。[1] Linton 和 Mammen 建立了参数部分的渐近理论以及模型非参数部分的逐点分布。

546

Su 和 Ullah(2006)考虑了具有未知形式非线性误差的非参数回归模型。他们建议用非参数核方法估计误差的序列相关结构，然后修正序列相关得到回归函数的更有效估计。Peng 和 Yao(2004)考虑了具有无限方差的依赖误差的非参数回归。

18.2.3　一期提前预测

考虑一个简单的非参数回归模型：

$$Y_t=g(Y_{t-1},\cdots,Y_{t-q})+u_t \tag{18.24}$$

① 关于 ARCH 和 GARCH 模型的创造性工作可参考 Engle(1982)和 Bollerslev(1986)。

它是式(18.11)的一个特殊情况,其中 $X_t=(Y_{t-1},\cdots,Y_{t-q})$。假定我们使用这个模型,用大小为 $n-q$ 的滚动估计样本做一期提前预测。对 $t=1,\cdots,m$,我们通过下式来预测 Y_{n+t}

$$\hat{Y}_{n+t,1}\equiv\hat{E}(Y_{n+t}\mid X_{n+t})=\frac{\sum_{j=q+t}^{n+t-1}Y_jK_h(X_j-X_{n+t})}{\sum_{j=q+t}^{n+t-1}K_h(X_j-X_{n+t})}\tag{18.25}$$

其中 $K_h(X_j-X_{n+t})=\prod_{s=1}^{q}h_s^{-1}k((Y_{j-s}-Y_{n+t-s})/h_s)$,因为 $X_j=(Y_{j-1},\cdots,Y_{j-q})$ 且 $X_{n+t}=(Y_{n+t-1},\cdots,Y_{n+t-q})$。

注意式(18.25)使用之前的 $n-q$ 个观测值计算 $E(Y_{n+t}|X_{n+t})$ 去预测 Y_{n+t}。在式(18.25)的右边,总和的上限是 $j=n+t-1$,所以,最近的 Y_j 是 Y_{n+t-1},它是预测期 $n+t$ 的前一期。因此,式(18.25)的右边只使用了到 $n+t-1$ 期之前的信息去预测 Y_{n+t}的值。这正好就是我们想要做的一期提前预测。在 m 期之后我们可以得到 m 个一期提前预测值 $\hat{Y}_{n+t}$,$t=1,\cdots,m$,且均方预测误差为

$$\text{MSFE}=\frac{1}{m}\sum_{t=1}^{m}(\hat{Y}_{n+t,1}-Y_{n+t})^2\tag{18.26}$$

在实践中,研究者也经常基于某个流行的参数模型去计算均方预测误差,547 比如 $Y_{n+t}=\alpha_0+\alpha_1Y_{n+t-1}+\cdots+\alpha_qY_{n+t-q}+u_t$。如果非参数均方预测误差(显著地)小于使用一个简单的参数模型得到的均方预测误差,这表示参数模型是设定错误的,那么应该寻找一个更加灵活的参数模型或者使用非参数模型去改进预测。

18.2.4 d 期提前预测

为了得出 d 期提前预测,我们可以考虑下面的回归模型:

$$Y_t=g(Y_{t-d},\cdots,Y_{t-d-q+1})+u_t\tag{18.27}$$

这也是式(18.11)的一个特殊形式,且 $X_t=(Y_{t-d},\cdots,Y_{t-d-q+1})$。假定我们使用这个模型,用大小为 $n-q-d$ 的滚动估计样本做 d 期提前预测。对 $t=1,\cdots,m$,我们预测 Y_{n+t}为

$$\hat{Y}_{n+t,d}\equiv\hat{E}(Y_{n+t}\mid X_{n+t})=\frac{\sum_{j=q+d-1+t}^{n+t-d}Y_jK_h(X_j-X_{n+t})}{\sum_{j=q+d-1+t}^{n+t-d}K_h(X_j-X_{n+t})}\tag{18.28}$$

其中 $K_h(X_j-X_{n+t})=\prod_{s=1}^{q}h_s^{-1}k((Y_{j+1-d-s}-Y_{n+t+1-d-s})/h_s)$。注意当 $d=1$ 时,式(18.28)应退化为式(18.25)。

注意式(18.25)使用之前 $n-q-d$ 个观测值计算 $E(Y_{n+t}|X_{n+t})$ 去预测 Y_{n+t}。在式(18.25)的右边,总和的上限是 $n+t-d$,所以,最近的 Y_j 是 Y_{n+t-d},它在预测期前 d 期被测量,即 $n+t$。因此,式(18.25)的右边只使用了到 $n+t-d$ 期之前的信息去预测 Y_{n+t}的值。对 m 期的提前 d 期均方预测误差为

$$\mathrm{MSFE}_d = \frac{1}{m}\sum_{t=1}^{m}(\hat{Y}_{n+t,d} - Y_{n+t})^2 \tag{18.29}$$

其中 $\hat{Y}_{n+t,d}$是在式(18.28)中被定义的。

在实践中,研究者经常基于某个流行的参数模型的均方预测误差来进行比较,比如 $Y_{n+t} = \alpha_0 + \alpha_1 Y_{n+t-d} + \cdots + \alpha_q Y_{n+t-d-q+1} + u_t$。

关于在预测框架中半参数神经网络模型的相关研究参见 Chen 和 White(1999)以及 Chen,Racine 和 Swanson(2001)。

18.2.5 非参数脉冲响应函数的估计

548

考虑下面的非参数条件异方差自回归模型:

$$Y_t = g(X_{t-1}) + \sigma(X_{t-1})u_t,\ t = q+1,\cdots,n \tag{18.30}$$

其中 $X_{t-1} = (Y_{t-1},\cdots,Y_{t-q})'$表示到滞后期 q 的滞后观测值向量,$g(\cdot)$和$\sigma(\cdot)$分别表示条件均值和条件标准离差,且 u_t 是具有零均值单位方差和有限第四阶矩的独立同分布序列。遵循 Koop,Pesaran 和 Potter(1996),Yang 和 Tschernig(2005)为水平 $k(\mathrm{GIR}_k)$定义了广义脉冲响应为这样的数量,即仅基于直到$t-1$期的信息,在时期 t 一个事先确定的冲击 u 改变 k 期提前预测的数量。也就是说,

$$\begin{aligned}\mathrm{GIR}_k(x,u) &= E(Y_{t+k-1} \mid X_{t-1} = x, u_t = u) - E(Y_{t+k-1} \mid X_{t-1} = x)\\ &= E(Y_{t+K-1} \mid Y_t = g(x) + \sigma(x)u, X_{t-1} = x) - E(Y_{t+k-1} \mid X_r\\ &\quad -1 = x)\end{aligned} \tag{18.31}$$

注意 GIR_k 依赖于 x 和冲击 u。一个非线性脉冲响应函数的替代定义由 Gallant,Rossi 和 Tauchen(1993)给出。Yang 和 Tschernig(2005)建议使用非参数核方法去估计包含在 GIR_k 中的多期提前预测。首先,我们引入一些记号。对 $k\geqslant 1$,我们记

$$g_k(x) = E[Y_{t+k-1} \mid X_{t-1} = x]$$

$$g_k(x,u) = E[Y_{t+k-1} \mid X_{t-1} = x, u_t = u]$$

且可得 $Y_{t+k-1} = g_k(X_{t-1}) + \sigma_k(X_{t-1})u_{t,k}$。

其中 $\sigma_k^2(x) = \mathrm{var}(Y_{t+k-1}|X_{t-1})$,且 $u_{t,k}$是鞅差分过程(因为 $E(u_{t,k}|X_{t-1}) = E(u_{t,k}|Y_{t-1},\cdots,Y_{t-q}) = 0$),其具有单位条件方差,即 $\mathrm{var}(u_{t,k}|X_{t-1}) = 1$。又令

$$\sigma_{k',k}(x) = \mathrm{cov}[(Y_{t+k'-1}, Y_{t+k-1}) \mid X_{t-1} = x]$$

$$\sigma_{k'k',k}(x) = \mathrm{cov}\{[(Y_{t+k'-1} - g_{k'}(X_{t-1}))^2, Y_{t+k-1} - g_k(X_{t-1})] | X_{t-1} = x\}$$

549　那么,我们可以把在式(18.31)中定义的 GIR_k 用更紧致的形式写为

$$\text{GIR}_k(x,u) = g_{k-1}(g(x)+\sigma(x)u,x) - g_k(x) = g_{k-1}(x_u) - g_k(x)$$

其中 $x_u=(g(x)+\sigma(x)u,x)$。用它们各自的局部线性核估计量替代多步预测函数,我们得到 GIR_k 的估计量为

$$\widehat{\text{GIR}}_k(x,u) = \hat{g}_{k-1}(\hat{x}_u) - \hat{g}_k(x)$$

其中 $\hat{x}=(\hat{g}(x)+\hat{\sigma}(x)u,x')$,我们在下面描述局部线性估计量 $\hat{g}(x)$、$\hat{g}_k(x)$ 和 $\hat{\sigma}(x)$。

定义 $e=(1,0_{1\times q})'$,$Z_k=\begin{pmatrix} 1 & \cdots & 1 \\ X_{q-1}-x & \cdots & X_{n-k}-x \end{pmatrix}'$

$$W_k = n^{-1}\text{diag}\{K_h(X_{t-1},x)\}_{i=q}^{n-k+1}$$

且 $Y_k=(Y_{q+k-1},\cdots,Y_n)'$。那么,$g_k(x)$ 的局部线性估计量为

$$\hat{g}_k(x) = e'(Z_k'W_kZ_k)^{-1}Z_k'W_kY_k \tag{18.32}$$

而

$$\hat{\sigma}_k(x) = \{e'(Z_k'W_kZ_k)^{-1}Z_k'W_kY_k^2 - \hat{g}_k(x)^2\}^{1/2} \tag{18.33}$$

注意 $\hat{g}(x)=\hat{g}_1(x)$ 且 $\hat{\sigma}(x)=\hat{\sigma}_1(x)$。在一些假定下,即 $h_s=cn^{-1/(q+4)}$,$c>0$,且 X_t 是一个平稳的 β 混合过程,且对某些 $c_0>0$ 和 $0<\rho<1$,其混合系数满足 $\beta(\tau)\leqslant c_0\rho^{-\tau}$,以及关于 $g(\cdot)$、σ 和 $f(\cdot)$(X_t 的密度)的附加的平滑性和矩条件,Yang 和 Tschernig(2005)推出了 $\widehat{\text{GIR}}_k$ 的渐近分布,即

$$\sqrt{nh_1\cdots h_q}\left\{\widehat{\text{GIR}}_k(x,u) - \text{GIR}_k(x,u) - \sum_{s=1}^{q} b_{g,k,s}(x,u)\right\} \xrightarrow{d} N(0,\boldsymbol{\Sigma}_k(x,u)$$

其中,偏误首项为

$$b_{g,k,s}(x,u) = b_{g,k-1,s}(x_u) - b_{g,k,s}(x) + \frac{\partial g_{k-1}(x_u)}{\partial x_1}[b_{g,s}(x) + b_{\sigma,s}(x)u]$$

550　其中 $b_{g,k,s}(x)=(\kappa_2/2)\partial^2 g_k(x)/\partial x_s^2$,且

$$b_{\sigma,k,s} = \kappa_2[\partial^2/\partial x_s^2(g_k^2(x)+\sigma_k^2(x)) - 2g_k(x)\partial^2/\partial x_s^2 g_k(x)]/[4\sigma_k(x)]$$

渐近方差为

$$\begin{aligned}\Sigma_{k(x,u)} = {} & \frac{\boldsymbol{\kappa}^q\sigma^2(x)}{f(x)}\left[\frac{\sigma_{k-1}^2(x_u)f(x)}{f(x_u)\sigma^2(x)} + \left(\frac{\partial g_{k-1}(x_u)}{\partial x_1}\right)^2\right. \\ & \left.\times\left\{1+u\mu_3+\frac{u^2(\mu^4-1)}{4}\right\}\frac{\sigma_k^2(x)}{\sigma^2(x)} - \frac{\partial g_{k-1}(x_u)}{\partial x_1}\left\{\frac{2\sigma_{1,k}(x)}{\sigma^2(x)} + \frac{u\sigma_{11,k}(x)}{\sigma^3(x)}\right\}\right] \\ & - \frac{\boldsymbol{\kappa}^q}{f(x)}\mathbf{1}(x=x_u)\left\{2\sigma_{k-1,k}(x) - 2\frac{\partial g_{k-1}(x_u)}{\partial x_1}\sigma_{1,k-1}(x) + u\frac{\partial g_{k-1}(x_u)}{\partial x_1}\frac{\sigma_{11,k-1}(x)}{\sigma(x)}\right\}\end{aligned}$$

其中 $\boldsymbol{\mu}_j=E(u_t^j)(j=3,4)$。

Yang 和 Tschernig(2005)也为 $\sigma_k^2(x)$ 提出了一个替代的即基于估计的非参

数残差项的估计量：

$$\tilde{\sigma}_k^2(x) = e'(Z'_k W_k Z_k)^{-1} Z'_k W_k V_k \tag{18.34}$$

其中 $V_k = ([Y_{q+k-1} - \hat{g}_k(X_{q-1})]^2, \cdots, [Y_n - \hat{g}_k(X_{n-k})]^2)'$。

上面描述的估计过程是很容易实施的，然而，因为累积了 k 个预测期的噪声，它可能会噪声过度。为了更有效地估计 $g_k(x)$，Yang 和 Tschernig(2005)建议了另一个多阶段方法，它是由 Chen，Yang 和 Hafner(2004)为估计 GIR_k 提出的，下面我们将简要介绍。令 $Y_t^{(0)} = Y_t$，且对 $j = 1, \cdots, k$ 重复下面的步骤，计算

$$\tilde{g}_j(x) = \frac{\sum_{t=m-1}^{n-k} K_h(X_t - x) Y_{t+j}^{(j-1)}}{\sum_{t=m-1}^{n-k} K_h(X_t - x)} \tag{18.35}$$

其中 $Y_{t+j}^{(j)} = \tilde{f}_j(X_t)$。最后一步 $j = k$ 给出 $\tilde{g}_k(x)$。Chen 等证明

$$\sqrt{nh_1 \cdots h_q}\left[\tilde{g}_k(x) - g_k(x) - \sum_{s=1}^{q} b_{g,k,s}(x) h_s^2\right] \xrightarrow{d} N(0, \kappa^q s_k(x)/f(x))$$

其中 $s_k(x) = \mathrm{var}(\tilde{f}_{k-1}(X_t) \mid X_{t-1} = x)$。由式(18.35)给出的多步预测量 $\tilde{f}_k(x)$ 和式(18.34)给出的条件标准离差估计量 $\tilde{\sigma}_k(x)(k = 1, \sigma(x) = \sigma_1(x))$，这推出了 GIR_k 的另一个估计量：

551

$$\widetilde{\mathrm{GIR}}_k(x, u) = \tilde{f}_{k-1}(\tilde{x}_u) - \tilde{g}_k(x)$$

Yang 和 Tschernig(2005)报告的模拟结果表明多过程估计量 $\widetilde{\mathrm{GIR}}_k$ 在均方误差意义上来说比直接的估计量 $\widehat{\mathrm{GIR}}_k$ 表现更好。

18.3　依赖数据的半参数模型

18.3.1　依赖数据的部分线性模型

Andrew(1994)考虑了一类一般的半参数模型，其包含半参数部分线性模型作为一个特例。我们在这一节考虑的具有依赖数据的部分线性模型为

$$Y_t = X_t'\gamma + \theta(Z_t) + u_t$$

我们考虑 β 混合模型的情形。因此，假定 (X_t', Z_t') 为具有混合系数 β_τ 的 β 混合过程，u_t 是一个鞅差分过程。我们使用与在独立数据情形中同样的方法估计 γ。也就是说，我们基于下面的回归方程估计 γ(见第 7 章的式(7.11))：

$$(Y_t - \hat{Y}_t)\hat{f}_t = (X_t - \hat{X}_t)'\hat{f}_t\gamma + \text{error}$$

其中 $\hat{X}_t = n^{-1}\sum_{s=1}^{n} X_s K_{h,ts}/\hat{f}_t$ 为 $\hat{E}(X_t \mid Z_t)$ 的核估计量，$\hat{Y}_t = n^{-1}\sum_{s=1}^{n} Y_s K_{h,ts}/\hat{f}_t$ 是 $\hat{E}(X_t \mid Z_t)$ 的核估计量，且 $\hat{f}_t = \hat{f}(Z_t)$。

正如第 7 章所述，γ 的半参数估计量是由 $(Y_t - \hat{Y}_t)\hat{f}_t$ 对 $(X_t - \hat{x}_t)\hat{f}_t$ 进行回归得到的最小二乘估计量，即

$$\hat{\gamma} = \left\{ \sum_{t=1}^{n} (X_t - \hat{X}_t)\hat{f}_t (X_t - \hat{X}_t)'\hat{f}_t \right\}^{-1} \sum_{t=1}^{n} (X_t - \hat{X}_t)\hat{f}_t (Y_t - \hat{Y}_t)\hat{f}_t \tag{18.36}$$

我们给出关于 β 混合过程的一个引理。

552 **引理 18.1** 假定 $X_1, \cdots, X_n$ 是一个随机向量($X_i \in \mathbb{R}^q$)的 β 混合过程。令 $g(x)$ 为 Borel 可测函数，且假定 $E|g_n(X_1)g_n(X_{1+\tau})| \leqslant M_n$，且对 $\delta > 0$ 和 $M_n > 0$，$E|g_n(X_1)|^{1+\delta} E|g_n(X_{1+\tau})|^{1+\delta} \leqslant M_n$，那么

$$|E[g_n(X_1)g_n(X_{1+\tau})] - E[g_n(X_1)]E[g_n(X_{1+\tau})]| \leqslant 4M_n^{1/(1+\delta)} |\beta(\tau)|^{\delta/(1+\delta)}$$

证明：见 Yoshihara(1976)的引理 1。

使用引理 18.1，Fan 和 Li(1999b)证明了下面的结论。

定理 18.5 在与定理 7.2 中给出的关于 $g(\cdot)$ 和 $f(\cdot)$ 的相似条件下，但是用 $\{Y_t, X_t, Z_t\}_{t=1}^{n}$ 是 β 混合过程，且其 β 混合系数 β_τ 满足对 $0 < \varepsilon < 1$ 和 $0 < \delta < \max\{1/2, 2/(q-2)\}$，$\beta_\tau^{\delta/(2+\delta)} = O(\tau^{-2+\varepsilon})$ 这一条件，替代独立同分布条件，那么

$$\sqrt{n}(\hat{\gamma} - \gamma) \to N(0, \Sigma) \text{ 依分布收敛}$$

其中

$$\begin{aligned}\Sigma &= \{E[v_t v_t' f^2(Z_t)]\}^{-1} E[\sigma^2(X_t, Z_t) v_t v_t' f^4(Z_t)] \{E[v_t v_t' f^2(Z_t)]\}^{-1} \\ &\equiv \Phi^{-1}\Omega\Phi^{-1}\end{aligned}$$

注意表达式 $\hat{\Phi} = n^{-1}\Sigma_t (X_t - \hat{X}_t)(X_t - \hat{X}_t)'\hat{f}_t^2$ 和 $\hat{\Omega} = n^{-1}\Sigma_t (\hat{u}_t \hat{f}_t)^2 (X_t - \hat{X}_t)\hat{f}_t (X_t - \hat{X}_t)'\hat{f}_t$ 分别是 Φ 和 Ω 的一致估计量，其中 $\hat{u}_t = (Y_t - \hat{Y}_t) - (X_t - \hat{x}_t)'\hat{\gamma}$。

比较定理 18.2 和 7.2，我们看到对弱依赖数据和独立数据，$\hat{\gamma}$ 的渐近分布是一样的。这不仅仅是因为弱依赖数据的假定，更重要的是因为我们假定 u_t 是一个鞅差分过程(因此是序列不相关的)。如果 u_t 为序列相关的，$\hat{\gamma}$ 的渐近方差可能有一个差分形式；参见 Newey 和 West(1978)，Andrews(1991b)，以及 Kiefer 和 Vogelsang(2002)关于异方差和自相关稳健协方差矩阵估计的最新研究。

18.3.2 可加回归模型

考虑一个如下形式的可加模型：

$$Y_t = g_1(X_{t1}) + \cdots + g_q(X_{tq}) + u_t \tag{18.37}$$

在弱依赖下的估计过程和我们在第 9 章中讨论的独立数据情形是一样的， 553
即我们可以用基于核的逆向拟合(backfitting)方法、边际积分方法或者非参数序列方法来估计式(18.37)。

令 $X_{ts} = Y_{t+1-d-s}$，$s = 1, \cdots, d$，那么式(18.37)成为

$$Y_t = g_1(Y_{t-d}) + \cdots + g_q(Y_{t-d-q+1}) + u_t \tag{18.38}$$

且如果我们使用核方法(比如，见 Kim 等(1999))，可以引入识别条件 $E[g_s(Y_{t-d-s})] = 0$，$s = 1, \cdots, q-1$，或者当使用序列(比如样条)方法时，要求 $g_s(0) = 0$，$s = 1, \cdots, q-1$。然后，模型(18.38)可以被用于做 d 期提前预测。

可加模型的一个缺点在于它们不允许在$(X_{t1}, \cdots, X_{tq})$中有一般的交互项。为了允许交互项同时保持非参数可加结构以避免维数诅咒，可以考虑我们在第 9 章中讨论的那种类型的部分线性可加模型。例如，当 $q = 2$ 且 $X_{ts} = Y_{t-d-s}$($s = 0, 1$)时，可以考虑以下模型：

$$Y_t = Z_t\alpha + g_1(Y_{t-d}) + g_2(Y_{t-d-1}) + u_t \tag{18.39}$$

其中 $Z_t = Y_{t-d}Y_{t-d-1}$。

这个模型的估计方法与我们在第 9 章的讨论是一样的。模型(18.39)允许交互项作为参数成分进入模型。

Gao 和 Tong(2004)考虑了在一般的具有依赖数据的部分线性模型中的模型选择问题。他们提出了一个在参数和非参数成分上选择滞后变量数目的一个新颖的交错鉴定过程。

18.3.3 依赖数据的变系数模型

在第 9 章考虑的变系数模型具有以下形式：

$$Y_t = X'_t\beta(Z_t) + u_t \tag{18.40}$$

其中 $X_t = (X_{t1}, \cdots, X_{tp})'$是 p 维的，$\beta(z) = (\beta_1(z), \cdots, \beta_p(z))'$是一个 $p \times 1$ 未知平滑函数向量，且 $Z_t = (Z_{t1}, \cdots, Z_{tq})$是 q 维的。

其估计方法和独立数据情形时一样。可以使用局部常数或者局部线性方 554
法来估计 $\beta(z)$。例如，$\beta(z)$的局部常数估计量为

$$\hat{\beta}(z) = \left[\sum_{j=1}^{n} X_j X'_j K_{h,jz}\right]^{-1} \sum_{j=1}^{n} X_j Y_j K_{h,jz} \tag{18.41}$$

其中 $K_{h,jz} = \prod_{s=1}^{q} h_s^{-1} k((Z_{js} - Z_s)/h_s)$ 为乘积核。

对于$\beta(z)$的局部线性估计量及其导函数估计量，在许多混合条件下，Cai，Fan 和 Yao(2000)，Cai，Fan 和 Li(2000)，以及 Fan，Yao 和 Cai(2003)证明了估计方法以及局部线性估计量的渐近分布和第 9 章讨论的独立数据情形是一样的。

在式(18.40)中，我们可得$X_t=(Y_{t-d},\cdots,Y_{t-d-p+1})$，尽管$Z_t$也是一个变量向量。然而，在实践中研究者经常选择$Z_t$为一个标量，有以下两个原因：首先，可以避免维数诅咒问题；其次，为了直观检查$\beta_s(z)$(可能是非线性)的函数形式，对$s=1,\cdots,p$，我们可以简单地画出z—$\beta_s(z)$的二维图表($s=1,\cdots,p$)。

18.4 半参数模型中的序列相关性检验

18.4.1 检验统计量及其渐近分布

我们再次讨论在 18.3.1 节中提出的半参数部分线性模型。回忆这些模型的形式为

$$Y_t = X'_t\gamma + \theta(Z_t) + u_t,\ t = -L+1,\cdots,0,1,\cdots,n \qquad (18.42)$$

其中X_t是$d\times 1$维，Z_t是$q\times 1$维，γ是$d\times 1$维参数，且$\theta(\cdot)$为未知平滑函数。我们所关注的假定是误差项u_t是否为零一阶序列相关或者零有限阶序列相关。

正如第 7 章中讨论的那样，为了检验u_t的零(有限阶)序列相关，使用二阶段法首先用下式估计γ，

$$\hat{\gamma} = S^{-1}_{(x-\hat{x})\hat{f}} S_{(x-\hat{x})\hat{f},(y-\hat{y})\hat{f}}$$

555

对于标量或者列向量数列，其第t个元素为$A_t\hat{f}_t$和$B_t\hat{f}_t$，我们使用记号$S_{A\hat{f},B\hat{f}}=\frac{1}{n}\sum_t A_t\hat{f}_t B'_t\hat{f}_t$和$S_{A\hat{f}}=S_{A\hat{f},A\hat{f}}$，且令$\hat{X}_t=\hat{E}(X_t|Z_t)$和$\hat{Y}_t=\hat{E}(Y_t|Z_t)$分别表示$E(X_t|Z_t)$和$E(Y_t|Z_t)$的核估计量。用$\hat{\gamma}$我们可以通过下式估计$u_t$：

$$\tilde{u}_t = Y_t - \hat{Y}_t - (X_t - \hat{X}_t)'\hat{\gamma} \qquad (18.43)$$

18.4.2 检验零一阶序列相关性

Li 和 Stengos(2003)建议用一个密度—加权检验统计量来检验缺乏零一阶序列相关性。他们在一个比假定误差u_t为鞅差分过程更强的假定下推导出其检验统计量的渐近分布。在这个条件下，我们有$\rho_f=E(u_t f_t u_{t-1} f_{t-1})=0$且$\rho=E(u_t u_{t-1})=0$。为了检验零一阶序列相关性，检验统计量的构造基于$\rho_f$或者$\rho$的样本对应值。Li 和 Stengos 构建了一个基于ρ_f的样本对应值检验统计量，因为它避免了在核估计量中出现一个随机分母所带来的技术困难。检验统计是基于

$$J_n \stackrel{\text{def}}{=} \frac{1}{\sqrt{n}} \sum_t \tilde{u}_t \tilde{u}_{t-1} \hat{f}_t \hat{f}_{t-1} \tag{18.44}$$

Li 和 Stengos(2003)证明了下面的结论。

定理 18.6　假定误差 u_t 是一个鞅差分过程，那么

(i) $J_n \to N(0,\sigma^2)$ 依分布收敛

其中 $\sigma_0^2 = E[\{(u_{t-1}f_{t-1} - v_t'f_tB^{-1}\Phi)u_tf_t\}^2]$，$\Phi = E[v_tf_tu_{t-1}f_{t-1}]$，$v_t = X_t - E(X_t|Z_t)$，$B = E[(X_t - \bar{X}_t)(X_t - \bar{X}_t)'(\hat{f}_t)^2]$。

(ii) $\hat{\sigma}_0^2 = \frac{1}{n}\sum_t \{[\tilde{u}_{t-1}\hat{f}_{t-1} - (X_t - \hat{x}_t)'\hat{f}_t\hat{B}^{-1}\hat{\Phi}]\tilde{u}_t\hat{f}_t\}^2$ 是 σ_0^2 的一致估计量，其中 $\hat{\Phi} = \frac{1}{n}\sum_t (X_t - \hat{x}_t)\hat{f}_t\tilde{u}_{t-1}\hat{f}_{t-1}$，$\hat{B} = \frac{1}{n}\sum_t (X_t - \hat{X}_t)(X_t - \hat{X}_t)'(\hat{f}_t)^2$

因此

$$J_n/\hat{\sigma}_0 \xrightarrow{d} N(0,1)$$

注意当 X_t 和 Z_t 是严格外生时，J_n 统计量是与 Durbin 和 Watson(1950)的检验统计量相似的，然而当 Z_t 包含了 Y_t 的滞后量，即 Y_{t-1} 时，它类似于 Durbin(1970)　556 的 h 统计量。Li 和 Stengos(2003)也考虑了检验更高阶的序列相关性。

上面对序列相关性的检验允许灵活的回归函数形式。然而，这样设计只是为了检查出特定形式的序列相关（比如，AR(1)），并不是检查未知形式的序列相关性的一致检验。Hong(1996)和 Pinkse(1988)提出了某些检验统计量对于参数回归函数形式假定中的任意类型的误差序列相关性都是一致的。

Robinson(1989)介绍了一个对经济时间序列的半参数和参数模型的一般的检验过程，并且将其应用到理性预期"惊奇模型"和市场非均衡模型。他的方法包括根据一个条件矩限制做出原假定。我们也可以使用 Robinson 的方法去检验零有限阶序列相关性，或者，更为一般地，检验一个形式为几乎处处 $E[Y_t|X_t,Z_t] = X_t'\gamma + \theta(Z_t)$ 的部分线性模型的正确设定。

18.5　依赖数据的模型设定检验

在这一节我们考虑检验正确的参数回归函数形式的原假定，同时考虑出现弱依赖数据的非参数显著性检验。在第 12 章和第 13 章中考虑到的其他检验都可以扩展到弱依赖数据。

18.5.1　对正确参数回归函数形式的一个核检验

为了叙述简便，我们关注检验线性原假定模型。我们考虑的原假定因此为

$$H_0^a: E(Y_t \mid X_t = x) = x'\gamma\text{，几乎所有 } x \tag{18.45}$$

至于独立数据情形,我们基于 $I^a = E[u_t E(u_t | X_t) f(X_t)]$ 构造我们的检验。I^a 的样本对应值为

$$\hat{I}_n^a = \frac{1}{n(n-1)} \sum_{t=1}^{n} \sum_{s \neq t}^{n} \hat{u}_t \hat{u}_s K_h (X_t - X_s)$$

557 其中 $\hat{u}_t = Y_t - X_t' \hat{\gamma}$,$\hat{\gamma}$ 是基于(零)线性模型的 γ 的最小二乘估计量。回忆我们使用 Hall(1984)中心极限定理去建立在假定 H_0^a 下的 $\hat{I}_n^a$ 的零分布。Hall 的中心极限定理已经由 Fan 和 Li(1999a),Hjellvik,Yao 和 Tjøstheim(1998)扩展到可以用于弱依赖(β 混合)数据。下面的定理由 Fan 和 Li 证明。

定理 18.7 假定 $h_s \to 0\ (s = 1, \cdots, q)$,当 $n \to \infty$ 时,$nH_q^{(7/8)} \to \infty$。同时假定 $f(X_t)$ 和 $E(Y_t | X_i)$ 的某些矩和平滑条件(详见 Fan 和 Li(1999a)),在 H_0^a 下可得

$$T_n^a \equiv n\,(h_1 \cdots h_q)^{1/2} \hat{I}_n^a / \hat{\sigma}_a \to N(0,1) \text{ 依分布收敛}$$

其中 $\hat{\sigma}_a^2 = \frac{2h_1 \cdots h_q}{n(n-1)} \sum_{t=1}^{n} \sum_{s \neq t}^{n} \hat{u}_t^2 \hat{u}_s^2 K_h^2 (X_t - X_s)$。

比较定理 18.7 和 12.1,可见不管数据是独立的还是弱依赖的,$\hat{T}_n^a$ 的渐近分布是一样的。

Fan 和 Li(1999a)也证明了如果 H_0^a 是非真的,T_n^a 以速率 $n\,(h_1 \cdots h_q)^{1/2}$ 趋于 $+\infty$。因此,T_n^a 是一个一致检验,即如果 H_0^a 为非真,当 $n \to \infty$ 时,拒绝零假定的概率趋于 1。Li(2005)建立了自举过程逼近 T_n^a 检验的零分布的有效性。另外,我们可以放松误差是序列无关的假定,而使用由 Berkowitz, Kilian 和 Birgean(1999)提出的重复抽样算法计算自举统计量,其对于未知形式的序列相关性是稳健的。

18.5.2 非参数显著性检验

Fan 和 Li(1996)以及 Lavergne 和 Vuong(2000)(在 12.3.4 节讨论的)考虑的非参数遗漏变量检验(即显著性检验)被 Li(1999)证明对于 β 混合数据是可行的。在误差项是鞅差分过程的假定下,检验统计量的构建和它们的渐近分布是和 12.3.4 节中给出的结果一样的,因此我们不再赘述。

对时间序列(即依赖)数据的模型设定检验有大量的文献。Juhl 和 Xiao

558 (2005)提出了一个对于在参数时间趋势模型中的结构变化的稳健的非参数检验(也参见 Delgado 和 Hidalgo(2000)对于在时间序列模型中发现结构突变的其他检验)。Kuan 和 Lee(2004)建议基于经验特征函数检验鞅差分假定。Chen, Chou 和 Kuan(2000)提出了一种方法检验时间可逆性,也是基于特征函数方法(这具有不用引入任何矩限制的优势,因为特征函数一般都是有界的),但是 Fan(1997)使用经验特征函数来检验一个多元分布函数的正确参数设定。

18.6　回归函数形式的非平滑性检验

我们也可以用一个非平滑检验来检验原假定 H_0^a：几乎处处 $E(Y_t \mid X_t) = m(X_t, \gamma)$。使用引理 A. 19（其包含平稳过程）给出的 Hilbert 空间中心极限定理是很容易做到这一点的。

Bierens 和 Ploberger（1997）考虑了基于下式的统计量：

$$\hat{J}_n(x) = n^{-1/2} \sum_{t=1}^{n} \hat{u}_t \mathcal{H}(X_t, x)$$

其中 $\hat{u}_t = Y_t - m(X_t, \hat{\gamma})$，且 $\hat{\gamma}$ 是 γ 的基于原模型的估计量（通常是$\sqrt{n}$一致的）。最终的条件矩（CM）检验为

$$\mathrm{CM}_n = \frac{1}{n} \sum_{j=1}^{n} [\hat{J}_n(X_j)]^2 \tag{18.46}$$

例如，可以使用$\mathcal{H}(X_t, x) = \mathbf{1}(X_t \leqslant x)$，或者在第 13 章中讨论的其他选择。Bierens 和 Ploberger（1997）证明 CM_n 在 H_0^a 下依分布收敛于 $\int [J_\infty(x)]^2 F(dx)$，$J_\infty(\cdot)$ 是定义良好的零均值高斯过程，且 $F(\cdot)$是 X_t 的累积分布函数。

Chen 和 Fan（1999）考虑了一个对于 β 混合数据的 CM_n类型非参数显著性检验，他们也证明可以使用一个平稳自举法（Politis 和 Romano（1994））去逼近其检验统计值的零分布。研究者也可以使用平稳自举法去逼近式（18.46）中定义的 CM_n 检验的零分布。

18.7　检验参数预测模型

559

18.7.1　条件累积分布函数的样本内检验

概率积分转换法

令 $F(y \mid \mathcal{I}_{t-1}) = F_t(y \mid \mathcal{I}_{t-1})$ 表示条件于直到 $t-1$ 期的信息集下 Y_t 的真实累积分布函数，令 $F(Y_t \mid \mathcal{I}_{t-1}, \theta_0) = F_t(Y_t \mid \mathcal{I}_{t-1}, \theta_0)$为参数条件累积分布函数，其中 θ_0 为一个有限维度参数。我们为了叙述方便，省略 $F(\cdot)$中的下标 t。我们要检验被选择的参数累积分布函数是否等于真实累积分布函数。于是，原假定为

$$H_0: P(Y_t \leqslant y \mid \mathcal{I}_{t-1}) = F(Y_t \mid \mathcal{I}_{t-1}, \theta_0) \text{ a. s.，某些 } \theta_0 \in \Theta \tag{18.47}$$

其中$\mathcal{I}_{t-1}$包含了直到 $t-1$ 期的所有相关变量，对立假定是对 H_0 的否定。

Diebold，Gunther 和 Tay（1998）建议检验 $U_t = F(Y_t \mid \mathcal{I}_{t-1}, \theta_0)$ 是否是均匀分布的。这是基于 $V_t = F(Y_t \mid \mathcal{I}_{t-1})$ 是独立同分布且在单位区间[0，1]内是均

匀分布的事实。因此,如果 $F(Y_t|\mathcal{I}_{t-1},\theta_0)$ 是正确设定的话,$F(Y_t|\mathcal{I}_{t-1},\theta_0)$ 在单位区间内也是均匀分布的。Diebold,Hahn 和 Tay(1999)将这种检验方法扩展到多变量 Y 情形,使得它可以用于评估密度预测的准确性,包括变量间的交互作用。Diebold 等(1998)检验的渐近分布是非标准的,且需要使用模拟或者自举的方法计算检验统计量的临界值。Bai(2003)提出使用一个鞅转换得到一个见系参数自由检验;也可参见 Pesaran 和 Timmermann(1992)的另一个预测检验。我们在下面简单地讨论 Bai 的检验方法。在 Markov 假定下,存在一个有限维变量 Z^{t-1},使得 $F(Y_t|Z^{t-1},\theta_0)=F(Y_t|\mathcal{I}_{t-1},\theta_0)$。令 $\hat{\theta}$ 表示一个基于原假定的 θ_0 的 $\sqrt{n}$ 一致估计量。定义 $\hat{U}_t=F(r|Z^{t-1},\hat{\theta})$,那么对于 $r\in[0,1]$,可以用 $n^{-1/2}\sum_{t=1}^{n}\mathbf{1}(\hat{U}_t\leqslant r)$ 估计 $F(r|Z^{t-1},\theta_0)$,且将它与真实条件累积分布函数 $F(r|Z^{t-1})=r$ 进行比较。于是,我们可以基于 $\hat{V}_n(r)\overset{\text{def}}{=}n^{-1/2}\sum_{t=1}^{n}[\mathbf{1}(\hat{U}_t\leqslant r)-r]$ 检验 H_0。Bai 证明

$$\hat{V}_n(r)=\frac{1}{\sqrt{n}}\sum_{t=1}^{n}[\mathbf{1}(U_t\leqslant r)-r]+\bar{g}(r)'\sqrt{n}(\hat{\theta}-\theta_0)+o_p(1) \quad (18.48)$$

560

其中

$$\bar{g}(r)=p\lim_{n\to\infty}n^{-1}\sum_{t=1}^{n}\frac{\partial F}{\partial\theta}(x|Z^{t-1},\theta_0)\big|_{x=F^{-1}(r|Z^{t-1},\theta_0)}$$

式(18.48)右边的第二项依赖于未知函数 $F(\cdot)$ 和 $(\hat{\theta}-\theta_0)$。因此,$\hat{V}_n(\cdot)$ 的渐近分布依赖于这些见系参数,且不能将这个检验的临界值列表显示。然而,通过应用一个鞅转换,我们可以移去 $\bar{g}^{(1)}(r)\sqrt{n}(\hat{\theta}-\theta_0)$ 这项,且这个转换过程会有一个见系参数的自由极限分布。令 $g^{(1)}$ 表示 $g(\cdot)$ 的导数,且定义 $C=\int_r^1 g^{(1)}(\tau)g^{(1)}(\tau)'\,\mathrm{d}\tau$。Bai(2003)建议了下面的见系参数分布自由检验统计量:

$$\hat{W}_n(r)=\hat{V}_n(r)-\int_0^r\left(g^{(1)}(s)'C^{-1}(s)\int_s^1 g^{(1)}(\tau)\,\mathrm{d}\hat{V}_n(\tau)\right)\mathrm{d}s \quad (18.49)$$

式(18.49)的右边第二项包含积分;Bai(2003,附录 B)提出了一个计算积分的数值方法。

在一些正则性条件下,Bai(2003,推论 1)证明了下面的结论:

$$H_0:\sup_{r\in[0,1]}|\hat{W}_n(r)|\xrightarrow{d}\sup_{r\in[0,1]}|W(r)| \quad (18.50)$$

其中 $W(r)$ 是一个标准布朗运动。于是,极限分布是没有见系参数的,且临界值可以列表表示,见 Bai。如果 H_0 非真,统计量 $\sup_{r\in[0,1]}|\hat{W}_n(r)|$ 以速率 $\sqrt{n}$ 趋近 $+\infty$,

从而它是一个一致检验。Bai 进一步讨论了怎样将他的检验应用到检验特定的参数条件累积分布函数,包括一般的 GRCH(p,q)过程和某些参数非线性时间序列模型。

Corradi 和 Swanson 的基于自举检验

Corradi 和 Swanson(2006)提出了一个对于参数条件累积分布函数正确设定的检验。Corradi 和 Swanson 的检验的一个不同特征是其允许动态误设。也就是说,Corradi 和 Swanson 假定变量 Z^{t-1}的选择集没有包含所有相关的历史信息。定义

$$\hat{U}_t = F(Y_t \mid Z^{t-1}, \tilde{\theta})$$

其中

561

$$\tilde{\theta} = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{t=1}^{n} \ln f(Y_t \mid Z^{t-1}, \theta)$$

且$f(\cdot \mid \cdot, \cdot)$是对应的参数条件概率密度函数。现在定义

$$\hat{V}_n(r) = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} [\mathbf{1}(\hat{U}_t \leqslant r) - r]$$

在某些合适的假定下,包括允许动态误设(Z^t 可能不等于 I_t),Corradi 和 Swanson(2006)证明在 H_0 假定下,

$$\hat{V}_n \stackrel{\text{def}}{=} \sup_{r \in [0,1]} | \hat{V}_n(r) | \stackrel{d}{\longrightarrow} \sup_{r \in [0,1]} | V(r) | \tag{18.51}$$

其中 $V(\cdot)$是一个具有零均值的高斯过程,且协方差函数依赖于来自参数估计和可能的动态误设的误差。

上面的 $\tilde{V}_n$ 不是没有见系参数的。Corradi 和 Swanson(2006)建议使用一个自举过程计算 $\tilde{V}_n$ 检验的临界值。这包括将再抽样序列的经验累积分布函数在自举估计量取值和真实序列的经验累积分布函数在基于真实数据的估计量取值之间进行比较。他们使用下面的重叠区块再抽样方案:在每个复制样本中,从原始样本$\{W_t\}_{t=1}^{n-1}$中的长度 l 中抽出 b 区块(有放回),其中 $W_t = (Y_t, Z^{t-1})$。例如,第一个区块是 $W_{i_1+1}, \cdots, W_{i_1+l}$,对某些 $i_1 \in \{1, \cdots, n-1-l\}$ 有概率 $1/(n-1-l)$,第二个区块是 $W_{i_2+1}, \cdots, W_{i_2+l}$,对某些 i_2 有概率 $1/(n-1-l)$(i_2 可以等于 i_1),等等。于是,再抽样序列 W_t^* 为 $W_1^*, \cdots, W_l^*, W_{l+1}^*, \cdots, W_n^* = W_{i_1+1}, \cdots, W_{i_1+l}, W_{i_2+1}, \cdots, W_{i_b+l}$。令 $\tilde{\theta}^*$ 表示使用自举样本得到的 θ 估计量,且定义

$$\hat{V}_n^*(r) = n^{-1/2} \sum_{t=1}^{n} (\mathbf{1}\{F(Y_t^* \mid Z^{*,t-1}, \tilde{\theta}^*) \leqslant r\} - \mathbf{1}\{F(Y_t \mid Z^{t-1}, \tilde{\theta}) \leqslant r\})$$

那么自举统计量为

$$\tilde{V}_n^* = \sup_{r \in [0,1]} | \tilde{V}_n^*(r) | \quad (18.52)$$

562 Corradi 和 Swanson(2006)建立了上面的自举过程的有效性。令 $\{\tilde{V}_{n,j}^*(r)\}_{j=1}^B$ 表示 B 自举统计量,且令 $\tilde{V}_{n,(\alpha)}^*$ 表示 $\{\tilde{V}_{n,j}^*(r)\}_{j=1}^B$ 的第 $1-\alpha$ 个百分位点。那么对于一个水平为 α 的检验,如果 $\hat{V}_n \geqslant \tilde{V}_{n,(\alpha)}^*$,拒绝假定 H_0;否则,不能拒绝 H_0。

18.7.2 条件累积分布函数的样本外检验

前一节是处理条件累积分布函数的正确设定的样本内检验。然而,如果对样本外预测有兴趣的话,那么对于评估样本外预测的精确度将会更加感兴趣。

对于样本外检验,需要将样本分成两部分,对估计样本有 R 个观测值,对预测样本有 P 个观测值($n=R+P$)。当进行预测时,模型是被递归地估计或者使用一个滚动样本来估计。Corradi 和 Swanson(2006)提供了使用两种估计方法的详细讨论。下面我们只讨论 Bai(2003)的检验以及 Corradi 和 Swanson 的检验的递归估计情形。

Bai(2003)的 V_n 统计量与之前的定义是一样的,即

$$\hat{V}_p(r) = \frac{1}{\sqrt{P}} \sum_{t=R}^{n-1} (\mathbf{1}\{F(Y_{t+1} | Z^t, \hat{\theta}_{t,\mathrm{rec}}) \leqslant r\} - r)$$

其中,$\hat{\theta}_{t,\mathrm{rec}}$ 是使用递归样本得到的 θ_0 的估计量,即

$$\hat{\theta}_{t,\mathrm{rec}} = \arg\min_{\theta \in \Theta} \frac{1}{t} \sum_{j=1}^{t} q(Y_j, Z^{j-1}, \theta),\ t = R, \cdots, n-1 \quad (18.53)$$

且 $q(Y_j, Z^{j-1}, \theta)$ 是目标函数,比如,$q(Y_j, Z^{j-1}, \theta) = -\ln f(Y_j, Z^{j-1}, \theta)$ 的拟极大似然估计,其中 $f(\cdot)$ 为给定 Z^{t-1} 时 Y_t 的(伪)密度。

对 $\hat{V}_p(r)$ 应用鞅变换得到

$$\hat{W}_{P,\mathrm{rec}(r)} = \hat{V}_P(r) - \int_0^r (g^{(1)}(s) C^{-1}(s) g^{(1)}(s)' \int_s^1 g(\tau) \mathrm{d}\hat{V}_p(\tau)) \mathrm{d}s$$

在 R 和 $P \to \infty$ 的假定下,$(\hat{\theta}_{t,\mathrm{rec}} - \theta_o) = O_p(P^{-1/2})$ 在 t 上均匀分布(这要求 563 $P/R \to \lambda, \lambda < \infty$),且与样本内检验情形中使用的类似的条件下,那么在假定 H_0 下

$$\sup_{\tau \in [0,1]} \hat{W}_{p,\mathrm{rec}}(r) \xrightarrow{d} \sup_{\tau \in [0,1]} W(r) \quad (18.54)$$

其中如前所述,$W(\cdot)$ 是一个标准布朗运动。

我们现在转向 Corradi 和 Swanson(2006)的样本外检验。首先,定义 $\hat{U}_{t+1,\mathrm{rec}} = F(Y_{t+1} | Z^t, \hat{\theta}_{t,\mathrm{rec}})$,且令 $V_{p,\mathrm{rec}} = P^{-1/2} \sum_{t=R}^{n-1} (\mathbf{1}(\hat{U}_{t+1,\mathrm{rec}} \leqslant r) - r)$

那么，在一些正则性条件下，包括 R 和 $P\to\infty$ 以及 $P/R\to\lambda, 0<\lambda<\infty$，Corradi 和 Swanson(2006)证明在假定 H_0 下，

$$\hat{V}_{p,\mathrm{rec}} \stackrel{\mathrm{def}}{=} \sup_{r\in[0,1]} \hat{V}_{p,\mathrm{rec}}(r) \xrightarrow{d} \sup_{r\in[0,1]} V(r) \tag{18.55}$$

其中 $V(\cdot)$ 是一个具有零均值和良好定义协方差函数的高斯过程。因为 V 不是没有冗余参数的，我们可以依赖再抽样过程去逼近它的零分布。

生成 $W_t^* = (Y_t^*, Z_t^{*,t-1}), t=1,\cdots,n$ 的区块自举法方案与我们先前讨论的样本内检验情形是完全一样的。再抽样序列由 b 个区块组成，它条件于样本，是离散的独立同分布均匀随机变量。定义 $\tilde{\theta}^*_{t,\mathrm{rec}}$ 为

$$\tilde{\theta}^*_{t,\mathrm{rec}} = \arg\min_{\theta\in\Theta} \frac{1}{t}\sum_{t=1}^{t}\Bigg(q(Y_j^*, Z^{*,j-1}, \theta) - \theta'\Bigg(\frac{1}{T}\sum_{k=1}^{n-1} \nabla \theta q(Y_k, Z^{k-1}, \hat{\theta}_{t,\mathrm{rec}})\Bigg)\Bigg) \tag{18.56}$$

其中 $R\leqslant t\leqslant n-1$。注意重要的是式(18.56)是围绕(总)样本均值波动(recentered)的，否则自举过程会有一个不可忽略的偏误项，即使是渐近的。

定义 564

$$V^*_{P,\mathrm{rec}}(r) = \frac{1}{\sqrt{P}}\sum_{t=R}^{n-1}\Bigg(\mathbf{1}[F(Y^{*,t}, \tilde{\theta}^*_{t,\mathrm{rec}}) \leqslant r] - \frac{1}{n}\sum_{j=1}^{n-1}\mathbf{1}[F(Y_{j+1} \mid Z^j, \hat{\theta}_{t,\mathrm{rec}} \leqslant r)]\Bigg) \tag{18.57}$$

自举统计量为

$$\hat{V}^*_{p,\mathrm{rec}} = \sup_{r\in[0,1]} \mid V^*_{P,\mathrm{rec}}(r) \mid \tag{18.58}$$

Corradi 和 Swanson(2006)证明 $\hat{V}^*_{p,\mathrm{rec}}$ 的自举分布提供了 $\hat{V}_{p,\mathrm{rec}}$ 零分布的一阶逼近。令 $\{\hat{V}^*_{p,\mathrm{rec},j}\}_{j=1}^{B}$ 表示自举统计量 B，且令 $\hat{V}^*_{p,\mathrm{rec},(\alpha)}$ 表示 $\{\hat{V}^*_{p,\mathrm{rec},j}\}_{j=1}^{B}$ 第 $1-\alpha$ 个百分位点，如果 $\hat{V}_{p,\mathrm{rec}} \geqslant \hat{V}^*_{p,\mathrm{rec},(\alpha)}$，拒绝 H_0；否则，不能拒绝 H_0。

18.8 应　用

18.8.1 预测短期利率

解释利率的期限结构在货币经济学中有丰富的历史；详细的讨论参见 Campbell(1995)和其中的参考文献。套利理论认为长期利率和短期利率之间必须是协整的(Campbell 和 Shiller(1987))。Anderson(1997)以及 Hanson 和 Seo(2000)使用参数非线性模型提供了相关证据，而 Bachmeier 和 Li(2002)使

用美国利率数据估计了一个半参数误差修正模型(ECM),发现了在期限结构中的非线性证据。也参见 White(2000)对以下统计问题的讨论:时间序列数据的重利用,以及对在一个特定的搜寻中遇到的最好的模型与给出的基准模型相比并没有预测优势这一原假定的检验过程。

Bachmeier 和 Li(2002)考虑了下面的基准线性 ECM,使用 BIC 来选择适当的滞后期。

$$\Delta r_t = \mu + \alpha \hat{W}_{t-1} + \gamma_1 \Delta R_{t-1} + \gamma_2 \Delta r_{t-1} + u_t \tag{18.59}$$

565

其中 R 是长期利率,r 是短期利率,$\hat{W}_{t-1} = R_{t-1} - \hat{\beta} r_{t-1}$ 是误差修正项,且 $\hat{\beta}$ 是估计的协整系数。

正如 Anderson(1997)以及 Hansen 和 Seo(2000)所主张的那样,线性性并没有隐含在期限结构理论中。Li 和 Wooldridge(2002)研究了这种可能性,一个半参数 ECM 可能提供一个对于潜在过程的更好的经验描述。Li 和 Wooldridge 考虑的半参数 ECM 为以下形式:

$$\Delta r_t = \theta(\hat{W}_{t-1}) + \gamma_1 \Delta R_{t-1} + \gamma_2 \Delta r_{t-1} + u_t \tag{18.60}$$

其中 $\theta(\cdot)$ 是一个未知函数形式。因此,式(18.60)是一个半参数部分线性模型。模型(18.60)与第 7 章中讨论的模型的不同点在于元素 $\hat{w}_{t-1}$ 是一个生成的回归元。Li 和 Wooldridge 提供了基于式(18.60)的估计和推断理论。式(18.59)和(18.60)的区别在于在式(18.59)中误差修正项是线性进入的,而在式(18.60)中误差修正项是非线性和非参数的。

遵循 Bachmeier 和 Li(2002),我们考虑在 McCulloch 和 Kwon(1993)中使用的期限结构数据。我们使用月度数据比较式(18.59)和(18.60)中的样本预测表现。我们使用 60 和 120 月利率(month rate)去预测从 1952 年到 1991 年间的 1、2、3、6、12 和 24 月利率。样本大小为 $n=480$。长期利率应包含短期利率的所有信息。

正如第 7 章讨论的,我们首先使用 $Y_t - \hat{E}(Y_t | W_t)$ 对 $X_t - \hat{E}(X_t | W_t)$ 进行回归来估计 $(\gamma_1, \gamma_2)'$,其中对 $s = 0, 1$, $\hat{E}(Y_{t-s} | W_{t-1}) = \sum_{j \neq t} Y_{j-s} K((\hat{W}_{t-1} - \hat{W}_{j-1})/h) / \sum_{j \neq t} K((\hat{W}_{t-1} - \hat{W}_{j-1})/h)$,且 $\hat{E}(X_{t-1} | W_{t-1}) = \sum_{j \neq t} X_{j-1} K((\hat{W}_{t-1} - \hat{W}_{j-1})/h) / \sum_{j \neq t} K((\hat{W}_{t-1} - \hat{W}_{j-1})/h)$,其中 $\hat{W}_{t-1} = Y_{t-1} - \hat{\beta} X_{t-1}$。令 $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2)'$ 表示 (γ_1, γ_2) 的最终的半参数估计量,Li 和 Wooldridge(2002) 证明 $\hat{\gamma}$ 是 $\sqrt{n}$ 一致的且具有渐近正态分布。我们用 $\hat{\theta}(W_{t-1}) = \hat{E}(Y_t | W_{t-1}) - \hat{\gamma}_1 \hat{E}(Y_{t-1} | W_{t-1}) - \hat{\gamma}_2 \hat{E}(X_{t-1} | W_{t-1})$ 来估计 $\theta(W_{t-1})$。

基于半参数误差修正模型的 Y_t 的预测值是 $\hat{Y}_{t,\text{semi}} = \hat{\theta}(W_{t-1}) + Y_{t-1}\hat{\gamma}_1 + X_{t-1}\hat{\gamma}_2$，且其均方预测误差是 $n_2^{-1}\sum_{t=n_1+1}^{n}(\hat{\gamma}_{t,\text{semi}} - Y_t)^2$，其中 $n_1 = 400$ 且 $n_2 = 80$（$n = n_1 + n_2 = 480$）。我们使用一个正态核，平滑参数是通过最小二乘交错鉴定法选择的。

为了进行比较，我们也计算由 $\hat{Y}_{t,\text{linear}} = \tilde{\mu} + \tilde{\alpha}\hat{W}_{t-1} + Y_{t-1}\tilde{\gamma}_1 + X_{t-1}\tilde{\gamma}_2$ 给出的 Y_t 的一期提前线性预测值，其中（$\tilde{\mu},\tilde{\alpha},\hat{\gamma}_1,\hat{\gamma}_2$）是（$\mu,\alpha,\gamma_1,\gamma_2$）的普通最小二乘估计量，使用的数据为 1 到 $t-1$ 期。这个线性模型的预测误差是 $Y_t - \hat{Y}_{t,\text{linear}}$ 且均方预测误差是 $n_2^{-1}\sum_{t=n_1+1}^{n}(\hat{Y}_{t,\text{linear}} - Y_t)^2$。

566

如表 18.1 所示，对于使用 10 年利率的半参数误差修正模型，一期提前预测均方误差总是比线性模型要好，因此半参数误差修正模型提供了一个改进的短期率的均方误差预测值。这说明在误差修正项中有非线性调整。

表 18.1 使用 10 年利率的平均预测均方误差

	1 月	2 月	3 月	6 月	12 月
线性	0.323	0.157	0.136	0.149	0.159
半参数	0.316	0.141	0.129	0.135	0.144

18.9 非平稳数据的非参数估计

我们所知道的唯一涉及非平稳数据的非参数渐近分析的文献有 Phillips 和 Park（1998），Juhl（2005），以及 Cai 和 Li（2006）。Phillips 和 Park 很显然是处理非平稳数据回归函数的非参数估计的第一篇文献，而 Juhl 考虑了非平稳数据的回归函数的半参数估计。Philips 和 Park 以及 Juhl 都考虑了一种情形，即真实 DGP 是一个线性单位根过程。

Cai 和 Li（2006）考虑了一个以下形式的变系数半参数回归模型：

$$Y_t = X_t'\beta(Z_t) + u_t \tag{18.61}$$

其中 Y_t 和 u_t 是标量变量，$X_t = (X_{t1},\cdots,X_{tp})'$ 是维度为 p 的协方差向量，且 $Z_t = (Z_{t1},\cdots,Z_{tq})$ 是一个维度为 q 的协方差（行）向量。Cai 和 Li 讨论了不同的情形，包括（i）X_t 是非平稳的，Z_t 是平稳的，（ii）X_t 是平稳的，Z_t 是非平稳的。Y_t 要么是平稳的，要么是非平稳的。考虑以下这种情形，即 X_t 是非平稳的，而 Z_t 和 Y_t 是平稳的。在这种情形下，模型（18.61）描述了 X_t 的不同部分之间的非线性协整关系。其估计方法与第 9 章中讨论的方法是一样的（见 9.3.1 节）。然而，渐近理论和独立数据或者平稳数据情形是有实质上不同的。对非平稳数据

567

非参数估计感兴趣的读者可以参考 Phillips 和 Park，Juhl，以及 Cai 和 Li 的进一步细节研究。

18.10 证 明

18.10.1 公式(18.9)的证明

证明：更多细节的计算表明 $\hat{f}(x)$ 的方差首项是和独立同分布数据情形一样的。使用（见习题 18.1）

$$\sum_{t=1}^{n-1}\sum_{s>t}^{n}\operatorname{cov}(K_{h,tx},K_{h,sx}) = \sum_{t=1}^{n-1}\sum_{j=1}^{n-t}\operatorname{cov}(K_{h,tx},K_{h,t+j,x})$$
$$= n\sum_{j=1}^{n-1}(1-j/n)\operatorname{cov}(K_{h,1x},K_{h,1+j,x})$$

可得 $(H_q = h_1,\cdots,h_q)$

$$\begin{aligned}
&\operatorname{var}((nH_q)^{1/2}\hat{f}(x))\\
&= (H_q/n)\left\{\sum_{t=1}^{n}\operatorname{var}(K_{h,tx}) + 2\sum_{t=1}^{n-1}\sum_{s>t}^{n}\operatorname{cov}(K_{h,tx},K_{h,sx})\right\}\\
&= (H_q/n)\left\{n\operatorname{var}(K_{h,tx}) + 2n\sum_{j=1}^{n}\left(1-\frac{j}{n}\right)\operatorname{cov}(K_{h,1x},K_{h,1+j,x})\right\}\\
&= H_q\operatorname{var}(K_{h,tx}) + 2H_q\sum_{j=1}^{n}\left(1-\frac{j}{n}\right)\operatorname{cov}(K_{h,1x},K_{h,1+j,x})\\
&= J_{1n} + J_{2n}
\end{aligned}\tag{18.62}$$

由式(18.5)可得

$$J_{1n} = \kappa^q f(x) + O(|h|^2)\tag{18.63}$$

568

其中 $\kappa = \int k(v_s)^2 dv_s$（因此 $\kappa = \int k(v)^2 dv$）。令 $\pi_n = [H_q^{-1/2}]$，是 $H_q^{-1/2}$ 的整数部分，且利用 $\sum_{j=1}^{n} = \sum_{j=1}^{\pi_n} + \sum_{j=\pi_n+1}^{n}$，我们可以把 J_{2n} 写为

$$\begin{aligned}
J_{2n} &= 2H_q\sum_{j=1}^{\pi_n}\left(1-\frac{j}{n}\right)\operatorname{cov}(K_{h,1x},K_{h,1+j,x})\\
&\quad + 2H_q\sum_{j=\pi_n+1}^{n}\left(1-\frac{j}{n}\right)\operatorname{cov}(K_{h,1x},K_{h,1+j,x})\\
&= J_{2n,1} + J_{2n,2}
\end{aligned}\tag{18.64}$$

令 $f_{1,1+j}(X_1,X_{1+j})$ 表示 (X_1,X_{1+j}) 的联合概率密度函数，可得

$|\operatorname{cov}(K_{h,1x},K_{h,1+j,x})| \leqslant |E[K_{h,1x},K_{h,1+j,x}]|$

$$= \left| \int K_h((x_1 - x)/h) K_h((x_{1+j} - x)/h) f_{1,1+j}(x_1, x_{1+j}) \mathrm{d}x_1 \mathrm{d}x_{1+j} \right|$$

$$= [f_{1,1+j}(x,x) \int K(u)K(v)\mathrm{d}u\mathrm{d}v + O(|h|)]$$

$$= [f_{1,1+j}(x,x) + O(|h|)] = O(1)$$

因此,

$$J_{2n,1} = H_q O(\pi_n) = O(\pi_n H_q) = o(1) \tag{18.65}$$

因为 $\pi_n = O(H_q^{-1/2})$。利用 ρ 混合的性质可得

$$\mathrm{cov}(K_{1x}, K_{1+j,x}) \leqslant \rho(j)\mathrm{var}(K_{1x}) = \rho(j) H_q^{-1}[f(x) \int K^2(v)\mathrm{d}v + O(|h|)]$$

因此,

$$J_{2n,2} \leqslant [f(x) \int K^2(v)\mathrm{d}v + o(1)] \sum_{j=\pi_n+1}^{\infty} \rho(j) \to 0 \tag{18.66}$$

由于 $\sum_{j=\pi_n+1}^{\infty} \rho(j) \to 0$,因为 $\pi_n \to \infty$ 。

结合式(18.62)和(18.66),我们已经证明

$$\mathrm{var}((nh_1 \cdots h_q)^{1/2} \hat{f}(x)) = \kappa^q f(x) + o(1)$$

在此完成了证明。 □

18.10.2 定理 18.2 的证明

569

证明:我们采用 Castellana 和 Leadbetter(1986)以及 Masry(1996a)的证明。记 $\hat{f}(x) - f(x) = \hat{f}(x) - E(\hat{f}(x)) + E(\hat{f}(x)) - f(x)$。我们已经证明 $\hat{f}(x)$ 的偏误首项和方差首项与独立数据情形下是一样的(见式(18.3)和(18.62))。为了证明渐近正态结果,我们首先引入某些表示法。

令 $Z_{n,t} = H_q^{1/2}[K_h(X_t, x) - E(K_h(X_t, x))]$,且令 $W_n = \sum_{t=1}^{n} Z_{n,t}$,那么

$$(nH_q)^{1/2}[\hat{f}(x) - E(\hat{f}(x))] = n^{-1/2} \sum_t Z_{n,t} = n^{-1/2} W_n$$

我们使用“大—小区块”理论。将集合 $\{1, \cdots, n\}$ 分成 $2k+1$ 个子集($k = k(n)$),其中大区块大小为 $a = a_n$,小区块大小为 $b = b_n$,其中 $k = k_n = [n/(a_n + b_n)]$([·]表示 · 的整数部分)。

我们引入当 $n \to \infty$ 时,$b_n \to \infty$ 且 $a_n \to \infty$,使得

$$b_n / a_n = o(1),\ a_n = o((nH_q)^{1/2}),\ \frac{n}{a_n} \rho(b_n) = o(1) \tag{18.67}$$

公式(18.67)只对 $\rho(\cdot)$ 施加了温和的条件。比如,我们可以选择 $a_n = [(nH_q)^{1/2} / \ln(n)]$,$b_n = a_n / \ln(n)$,所以,公式(18.67)的前两个条件成立。如果

用最优平滑法进一步选择 h_s，即 $h_s = O(n^{-1/(4+q)})$，那么，使得式(18.67)的第三个条件成立的充分条件是当 $\tau \to \infty$ 时，$\tau^{1+q/2}\rho(\tau) \to 0$，这个条件比用来证明收敛速度结果(定理18.1)的条件 $\rho(\tau) = O(\tau^{-(1+\varepsilon)})$ 稍强(对任意小的 $\varepsilon > 0$)。

定义随机变量

$$\eta_j = \sum_{i=j(a+b)}^{j(a+b)+a-1} Z_{n,i}, 0 \leqslant j \leqslant k-1 \tag{18.68}$$

$$\zeta_j = \sum_{i=j(a+b)+a}^{(j+1)(a+b)-1} Z_{n,i}, 0 \leqslant j \leqslant k-1 \tag{18.69}$$

$$\xi_k = \sum_{i=k(a+b)}^{n} Z_{n,i} \tag{18.70}$$

η_j 具有 a_n 的较大长度，ζ_j 具有 b_n 的较小长度，且 ξ_k 具有小于 a_n 的长度。

570

记

$$W_n = \sum_{j=0}^{k-1} \eta_j + \sum_{j=0}^{k-1} \zeta_j + \xi_k \equiv W_{1n} + W_{2n} + W_{3n}$$

显然，W_{1n} 是 W_n 的首项，因为 W_{1n} 包含 $k_n a_n$ 项，而 W_{2n} 包含 $k_n b_n$，W_{3n} 只包含 a_n 项。

下面我们将证明

$$n^{-1}E[W_{2n}^2] = o(1), n^{-1}E[W_{3n}^2] = o(1) \tag{18.71}$$

这意味着 $n^{-1/2}W_{2n} = o_p(1)$ 且 $n^{-1/2}W_{3n} = o_p(1)$。

对于 W_{1n}(注意 $\exp(itW_{1n}) = \prod_{j=0}^{k} \exp(it\eta_j)$)($i = \sqrt{-1}$)，我们将证明对每一个 $\varepsilon > 0$，

$$\left| E[\exp(itW_{1n})] - \prod_{j=0}^{k-1} E[\exp(it\eta_j)] \right| \to 0 \tag{18.72}$$

$$\frac{1}{n}\sum_{j=0}^{k-1} E[\eta_j^2] \to \Omega(x) \tag{18.73}$$

$$\frac{1}{n}\sum_{j=0}^{k-1} E[\eta_j^2 \mathbf{1}(|\eta_j| > \varepsilon\Omega(x)/\sqrt{n}] \to 0 \tag{18.74}$$

公式(18.72)说明 W_{1n} 的特征函数与 η_j 独立于 j 的过程是渐近相同的，这就意味着在 W_{1n} 内 $\{\eta_j\}$ 的加总是渐近独立的。公式(18.73)和(18.74)是关于 W_{1n} 在独立条件下的渐近正态性的 Lindeberg-Feller 条件。于是，式(18.72)和式(18.74)意味着 $\sqrt{n}W_{1n} \xrightarrow{d} N(0, \Omega(x))$。

我们现在证明这些结论。首先考虑 W_{2n}。

$$\begin{aligned} E[W_{2n}^2] &= \operatorname{var}\left[\sum_{j=0}^{k-1} \zeta_j\right] = \sum_{j=0}^{k-1} \operatorname{var}(\zeta_j) + \sum_{i=0}^{k-1}\sum_{j=0, j\neq i}^{k-1} \operatorname{cov}(\zeta_i, \zeta_j) \\ &\equiv F_{1n} + F_{2n} \end{aligned} \tag{18.75}$$

在 ζ_i 当中有 b_n 项，利用与推导式(18.62)同样的论证过程，可得　　571

$$\operatorname{var}(\zeta_i) = b_n \operatorname{var}(Z_{n,i}) + 2b_n \sum_{i=1}^{b_n-1} (1 - b_n^{-1}) \operatorname{cov}(Z_{n,1}, Z_{n,t})$$

$$= b_n \Omega(x)[1 + O(b_n H_q)] = b_n \Omega(x)[1 + o(1)] \quad (18.76)$$

因此，利用式(18.75)和式(18.76)可知

$$F_{1n} = k_n b_n \Omega(x)[1 + o(1)] = O(k_n b_n) = o(n) \quad (18.77)$$

因为 $k_n(a_n + b_n) = O(n)$ 和 $b_n = o(a_n)$ 意味着 $k_n b_n = o(n)$。

同样，可得

$$|F_{2n}| \leqslant 2 \sum_{i_1=0}^{n-a_n-1} \sum_{i_2=i_1+a_n}^{n-1} |\operatorname{cov}(Z_{n,i_1}, Z_{n,i_2})| \leqslant 2n \sum_{j=a_n}^{\infty} |\operatorname{cov}(Z_{n,1}, Z_{n,1+j})|$$

$$\leqslant 2n \operatorname{var}(Z_{n,1}) \sum_{j=a_n}^{n} \rho(j) = o(n) \quad (18.78)$$

因为 $\operatorname{var}(Z_{n,1}) = \Omega(x) + o(1) = O(1)$ 且 $\sum_{j=a_n}^{\infty} \rho(j) \to 0$ 因为 $a_n \to \infty$。

因此，$n^{-1}E[W_{2n}^2] = o(1)$，这意味着 $n^{-1/2}W_{2n} = o_p(1)$。

对于 W_{3n}，显然 W_{3n} 的阶数小于 W_{1n} 的阶数，因为 W_{1n} 关于 $a_n k_n$ 加总，而 W_{3n} 加总不多于 a_n 项。我们下面将证明 $\sqrt{n}W_{1n} = O_p(1)$，这意味着 $\sqrt{n}W_{3n} = o_p(1)$。我们将使用下面的关于 ρ 混合过程的不等式。

引理 18.2　令 $V_1, \cdots, V_J$ 为 α 混合随机变量，分别属于 $\mathcal{F}_{i_1}^{j_1}, \cdots, \mathcal{F}_{i_J}^{j_J}$，其中 $1 \leqslant i_1 \leqslant j_1 < i_2 < \cdots < j_J \leqslant n, i_{l+1} - j_l \geqslant \tau$，且 $|V_j| \leqslant 1, j = 1, \cdots, J$，那么

$$\left| E\left[\prod_{j=1}^{J} V_j\right] - \prod_{j=1}^{J} E[V_j] \right| \leqslant 16(J-1)\alpha(\tau)$$

证明：参见 Volkonskii 和 Rozanov (1959) 或者 Hall 和 Heyde (1980，定理 A.5)。　□

上面的结论对于 ρ、β 和 ϕ 混合过程都成立，因为 $\alpha(\tau)$ 不大于其他任意一个混合系数。　572

应用引理 18.2 于 $\exp(itW_n) = \prod_{j=0}^{k-1} \exp(it\eta_j)$ (其中 $V_j = \exp(it\eta_j)$)，由条件(18.67)可得

$$\left| E[\exp(itW_{1n})] - \sum_{j=0}^{k-1} E[\exp(it\eta_j)] \right| \leqslant Ck_n \rho(a_n) = O\left(\frac{n}{a_n}\right)\rho(b_n) = o(1)$$

我们可以选择 $a_n = (nH_q)^{1/2}/\ln(n)$，$b_n = (nH_q)^{1/2}/[\ln(n)]^2$，$h = O(n^{-1/(4+q)})$ (最优平滑)，所以 $O(\frac{n}{a_n})\rho(b_n) = o(1)$ 的一个充分条件是当 $\tau \to \infty$ 时，$\tau^{1+q/2}\rho(\tau) \to 0$。

由平稳性条件和推导式(18.76)同样的论证可得

$$\mathrm{var}(\eta_j) = \mathrm{var}(\eta_0) = a_n \Omega(x)[1 + o(1)]$$

因此,

$$\frac{1}{n}\sum_{j=0}^{k-1} E[\eta_j^2] = \frac{k_n a_n}{n}\Omega(x)[1 + o(1)] \to \Omega(x)$$

因为 $k_n a_n / n \to 1$(因为 $b_n / a_n \to 0$)。同样可得

$$|Z_{n,t}| = H_q^{-1/2}\left|K\left(\frac{X_t - x}{h}\right) - E\left[K\left(\frac{X_t - x}{h}\right)\right]\right| \leqslant CH_q^{-1/2}$$

因为 $K(\cdot)$是有界的。这意味着由式(18.68)可得

$$\max_{0 \leqslant j < k-1} |\eta_j| / \sqrt{n} \leqslant Ca_n / (nH_q)^{1/2} \to 0 \tag{18.79}$$

因此当 n 很大时,式(18.79)意味着集合$\{|\eta_j| \geqslant \varepsilon \Omega(x)\sqrt{n}\}$为一个空集,于是式(18.74)成立。在此完成了定理 18.2 的证明。 □

18.11 习 题

习题 18.1 证明

$$\sum_{t=1}^{n-1}\sum_{j=1}^{n-t} \mathrm{cov}(K_{h,tx}, K_{h,t+j,x}) = n\sum_{j=1}^{n-1}(1 - j/n)\mathrm{cov}(K_{h,1x}, K_{h,1+j,x})$$

573

习题 18.2 证明式(18.14),即证明

$$\mathrm{var}[\hat{m}_1(x)] = O(\|h\|^2 (nH_q)^{-1})$$

习题 18.3 使用引理 18.1,证明

$$\hat{f}(x) - f(x) = O_p(\|h\|^2 + (nH_q)^{-1/2})$$

对于一个平稳 β 混合过程,其混合系数满足

$$\sum_{j=1}^{\infty} \beta(j)^{\delta/(1+\delta)} < \infty$$

提示:偏误的计算和独立数据情形是一样的。用引理 18.1,可得

$$|\mathrm{cov}(K_{h,tx}, K_{h,sx})| = |E\{[K_{h,tx} - E(K_{h,tx})][K_{h,sx} - E(K_{h,sx})]\}| \leqslant 0 + 4M_n^{1/(1+\delta)}$$

可以进一步证明

$$M_n \sim E[(K_{h,tx}K_{h,sx})^{1+\delta}] = \int (K_{h,tx}K_{h,sx})^{1+\delta} f(x_t, x_s)\,\mathrm{d}x_t \mathrm{d}x_s \leqslant CH_q^{-2\delta}$$

那么,

$$\begin{aligned}\left|\sum_t \sum_{s \neq t} \mathrm{cov}(K_{h,tx}, K_{h,sx})\right| &\leqslant 4M_n^{1/(1+\delta)} \sum_t \sum_{s>t} \beta(s-t)^{\delta/(1+\delta)} \\ &\leqslant 4M_n^{1/(1+\delta)} n \sum_{j=1}^{\infty} \beta(j)^{\delta/(1+\delta)} \\ &= O(nH_q^{-2\delta/(1+\delta)}) = o(2H_q^{-1})\end{aligned}$$

因为 $\delta < 1$。剩下的步骤显而易见。

第19章　面板数据模型

575

"面板"是通过在 T 个连续时期对 N 个横截面单元抽取观测值所形成的一个样本，得到形如 $\{Y_{it},Z_{it}\}_{i=1,t=1}^{N,T}$ 的一个数据集。因此一个面板简单地就是 N 个个体时间序列的一个结合，时间序列可能很短（"小的 T"）或者很长（"大的 T"）。例如，美国收入动态面板研究（Panel Study of Income Dynamics, PSID）[①]是一个全国代表性纵向研究，包括从1968年开始的超过7 000户的美国家庭，而加拿大的劳动和收入动态调查（Survey of Labonr and Income Dynamics, SLID）[②]是一个由加拿大统计局实施的纵向家庭调查，统计局跟踪了大约15 000户加拿大代表性家庭样本达6年之久，最早的面板开始于1993年。同美国相对应的，SLID面板被设计用来捕捉随着时间的推移个体和家庭经济福利的变化以及福利的决定因素。

对面板数据进行建模时，会产生许多有趣的统计问题，包括如何模型化横截面效应（即是否考虑不可观测的异方差性问题），是否汇总数据，以及如何模型化潜在的依赖等。已经存在很多关于面板数据模型的优秀文献，我们建议感兴趣的读者参见Arellano(2003)的相关文章进行全面的了解。在本章，我们考虑许多已经被提出来的对面板数据建模的灵活方法。

我们将关注两种类型的半参数和非参数模型：(i)部分线性半参数面板数据模型和完全非参数面板数据模型，(ii)具有未知误差分布的面板离散选择模型和

576

① 参见 http://psidonline.isr.umich.edu/。

② 参见 http://www.statcan.ca/English/sdda/3889.htm。

审查回归模型。我们将泄盖随机效应和固定效应的非参数面板数据模型。

在本章,我们令下标 $i=1,2,\cdots,N$ 代表个人, $t=1,2,\cdots,T$ 表示时间。通常,T 相对于 N 是较小的。为了论述的清晰,我们将讨论限定于平衡面板(即 T 没有随着 i 而改变)。但是,我们得到的结果很容易扩展到非平衡面板的情况。

我们也将在 19.8 节看到一个如下形式的部分线性模型能够很自然地出现在面板数据设定中:

$$Y_{it} = X_{it}'\alpha + g(Z_{it}) + u_{it} \tag{19.1}$$

对于面板数据模型一个流行的误差设定是所谓的单向误差成分模型(one-way error component model),其误差过程给定为 $u_{it}=\mu_i+\nu_{it}$,其中 μ_i 是(随机或固定的)个体效应,ν_{it}是异质误差项(idiosyncratic error term)。ν_{it}通常被假定为关于下标 i 和 t 都成立的一个独立过程。对于随机效应模型,个体效应 μ_i 的存在使得误差项 u_{it}序列相关。在这种情况下,若 Z_{it}包括一个滞后的因变量,如 $Y_{i,t-1}$,则 $E(u_{it}|Z_{it})\neq 0$,并且为了估计式(19.1)得采用一些工具变量(IV)法(参见第 16 章对于非参数和半参数工具变量方法的一个综述)。然而,若式(19.1)中的误差 u_{it}是序列不相关的,例如,若 u_{it}是一个鞅差分过程(它排除了一个误差成分模型),接着则可以如第 18 章讨论过的那样$\left(将 \sum_t 替换为 \sum_i\sum_t\right)$估计式(19.1)。因此,在实践中,研究者可能想要首先检验零序列相关或者是检验没有个体效应的存在,而后才决定误差成分模型是否适合于手头上的数据。

19.1 面板数据模型的非参数估计:忽略方差结构

一般的非参数面板数据模型给定为:

$$Y_{it} = g(Z_{it}) + u_{it}, \quad i=1,2,\cdots,N,\ t=1,2,\cdots,T \tag{19.2}$$

其中,为了表述的简洁,我们仅仅考虑 Z_{it}是一个 q 维的连续随机向量的情况。

577 我们假定数据关于下标 i 是独立的。若 N 较大,T 较小,则没必要假定数据关于 t 是平稳的。然而,若 T 很大,则通常需要假定数据是平稳的,并且服从某种类型的混合过程(参见第 18 章对于各种混合过程的定义)。

一个标准的方法是引入一个个体效应 μ_i。我们首先考虑随机个体效应的情况;然而,在本节,我们将保留 Z_{it}是严格外生的假定,即对所有 t,$E(u_{it}|Z_{i1},Z_{i2},\cdots,Z_{iT})=0$。在这个假定下,可以使用一个标准的非参数方法来估计 $g(\cdot)$,比如局部常数法,即

$$\hat{g}(z) = \frac{\sum_{j=1}^{N}\sum_{s=1}^{T} Y_{js} K_{h,js,z}}{\sum_{j=1}^{N}\sum_{s=1}^{T} K_{h,js,z}} \quad (19.3)$$

其中$K_{h,js,z} = \prod_{l=1}^{q} h_l^{-1} k((Z_{js,l} - Z_l)/h_l)$。

如果对于一个固定的 i,u_{it}是一个满足在第 18 章中给定的条件的混合过程,则 $\hat{g}(x)$的渐近分布与独立数据情形下是一样的。下面我们考虑一个单向误差成分模型的情形

$$u_{it} = \mu_i + \nu_{it}$$

其中 μ_i 是独立同分布$(0,\sigma_{\mu}^2)$,ν_{it}是一个零均值有限方差 β 混合过程。我们也假定 X_{it}是一个(平稳)混合过程,其混合系数满足 $\sum_{\tau=1}^{\infty} \beta_{\tau}^{\delta/(1+\delta)} < \infty$。那么,在类似于我们第 2 章 2.1 节介绍过的施加于 $g(\cdot)$和$f(\cdot)$上的平滑性条件下,我们有下面的结论:

定理 19.1 假定当 $NT \to \infty$ 时,$NTH_q \to \infty$ $(H_q = (h_1 \cdots h_q))$。同样假定如下的条件之一成立:(i) N 较大,T 较小;(ii) N 和 T 都较大,且当 $T \to \infty$ 时,$TH_q \to 0$,那么

$$\sqrt{NTH_q}\left[\hat{g}(z) - g(z) - \sum_{s=1}^{q} h_s^2 B_s(z)\right] \xrightarrow{d} N\left(0, \frac{\kappa^q \sigma^2(z)}{f(z)}\right) \quad (19.4)$$

其中 $B_s(z)$与 2.1 节中定义的相同。

定理 19.1 的证明在 19.10 节中给出。

注意当 T 很小时,没必要对(Z_{it},ν_{it})施加混合条件。当 T 很大时,我们做出额外的假定,当 $T \to \infty$ 时,$TH_q \to 0$。这个条件被用来控制协方差 $\mathrm{var}(\hat{g}(z))$(因为 μ_i 的存在)。

定理 19.1 没有包含 N 较小、T 较大的情形,因为单向(随机效应)误差成分模型在这种设定下不是一个合理的方法。当 T 较大、N 较小时,可以不用汇合(pool)数据,因为对于每一个个体单元都存在一个较长的时间序列。在这种情况下,对每个个体单元 i 可以使用对每个个体单元可用的 T 个个体时间序列来简单地估计 $g_i(Z_{it})$,最终估计量的渐近分布在第 18 章中讨论过。为了允许数据实际上是可汇合的这种可能性,可以引入一个无序离散变量,比如对 $i=1,2,\cdots,N$,$\delta_i = i$,使用第 4 章介绍过的混合离散和连续核方法来非参数地估计 $E(Y_{it}|Z_{it},\delta_i) = g(Z_{it},\delta_i)$。令 $\hat{\lambda}$ 表示与 δ_i 有关的交错鉴定平滑参数,则若 $\hat{\lambda} = 1$,得到 $g(Z_{it},\delta_i) = g(Z_{it})$,并且因此数据在 $g(\cdot)$的最终估计中是可汇合的。 578

另一方面,若 $\hat{\lambda}=0$(或者接近于 0),则对第 i 个个体单元仅仅使用时间序列就可以有效地估计每个 $g_i(\cdot)$。最后,若 $0<\hat{\lambda}<1$,可以解释为数据是部分可汇合的情况。

我们也可以应用 Baltagi, Hidalgo 和 Li(1996)以及 Lavergne(2001)提出的非参数可汇合性检验来检验数据是否是可汇合的。

19.2 Wang的有效非参数面板数据估计量

Wang(2003)考虑了如下的非参数面板数据回归模型:

$$Y_{it} = g(Z_{it}) + u_{it},\quad i = 1,2,\cdots,N,\ t = 1,2,\cdots,T \qquad (19.5)$$

其中 $g(\cdot)$是一个未知的平滑函数, $Z_{it}=(Z_{it,1},\cdots,Z_{it,q})$是 q 维的,所有其他变量都是标量,并且 $E(u_{it}|Z_{i1},\cdots,Z_{iT})=0$。

在这一节,我们将仅仅考虑 N 较大、T 较小的情况,并且我们假定数据关于 i 是独立的,关于 t 是严格平稳的。接下来,我们令 $u_i=(u_{i1},\cdots,u_{iT})'$, $\Sigma=\text{cov}(u_i|Z_{i1},\cdots,Z_{iT})$。Wang(2003)提出了一种新的方法来估计 $g(\cdot)$,这种方法利用了包含在 Σ 中的信息。我们将首先讨论对于已知 Σ 的估计方法,然后讨论当 u_{it}是单向误差成分结构时估计 Σ 的情形。

579 Wang(2003)建议对于个体 i 通过式(19.6)来选择准则函数

$$\mathcal{L}_i(\cdot) = \mathcal{L}(Y_i,g_i) = -\frac{1}{2}[Y_i - g_i]'\Sigma^{-1}[Y_i - g_i] \qquad (19.6)$$

其中 $Y_i=(Y_{i1},\cdots,Y_{iT})'$, $g_i=(g_{i1},\cdots,g_{iT})'$, $g_{it}=g(Z_{it})$。

我们令$\mathcal{L}_{i,tg}=\partial\mathcal{L}_i(\cdot)/\partial g_{it}$, $\mathcal{L}_{i,tsg}=\partial^2\mathcal{L}_i(\cdot)/(\partial g_{it}\partial g_{is})$。从式(19.6)中我们得到

$$\mathcal{L}_{i,tg} = c_t'\Sigma^{-1}[Y_i - g_i] = \sum_{s=1}^{T}\sigma^{ts}(Y_{is} - g(Z_{is})) \qquad (19.7)$$

其中 c_t 是一个第 t 个元素等于 1、所有其他元素等于 0 的 $T\times 1$ 维向量,其中 σ^{ts} 是 Σ^{-1}的第(t,s)个元素。同样,我们有

$$\mathcal{L}_{i,tsg} = -c_t'\Sigma^{-1}c_s = -\sigma^{ts} \qquad (19.8)$$

我们将最大化一个加权核的目标函数。定义乘积核

$$K_h(v) = \prod_{j=1}^{q} h_j^{-1}k(v_j/h_j)$$

其中 $k(\cdot)$是一个一元核函数。令 $f(\cdot)$表示 Z_{it}的概率密度函数。定义 $G_{it}(z,h)=(1,(Z_{it,1}-Z_1)/h_1,\cdots,(Z_{it,q}-Z_q)/h_q)'$(注意 $G_{it}(z,h)$是$(q+1)\times 1$ 维),定义 $g^{(1)}(z)=\partial g(z)/\partial z$是 $g(\cdot)$关于 z 的一阶偏导数(是 $q\times 1$ 维)。Wang

(2003)建议通过 $\alpha_0(z)$ 来估计未知函数 $g(z)$，其中 $\alpha_0(z)$ 和 $\alpha_1(z)$ 是以下一阶条件的解

$$0=\sum_{i=1}^{N}\sum_{t=1}^{T}K_h(Z_{it},z)G_{it}(z,h)\,\mathcal{L}_{tg}(Y_i,\hat{g}(Z_{i1}),\cdots,\alpha_0(z)+[(Z_{it}-z)/h]'\alpha_1(z),\cdots,\hat{g}(Z_{iT}))\tag{19.9}$$

其中 $\hat{g}(Z_{is})$，$s\neq t$ 是 $g(Z_{is})$ 的初始估计量，且 $[(Z_{it}-z)/h)]'=((Z_{it,1}-Z_1)/h_1,\cdots,(Z_{it,q}-Z_q)/h_q)$。注意式(19.9)是一个 $(q+1)\times 1$ 维的方程，从中我们求解得 $(q+1)$ 个参数 $\hat{\alpha}_0(z)$（一个标量）及 $\hat{\alpha}_1(z)$（一个 $q\times 1$ 向量），它们分别估计了 $g(z)$ 和 $g^{(1)}(z)$。

一个迭代过程

方程(19.9)暗示了如下的迭代估计过程。假定在第 $[l-1]$ 步 $g(\cdot)$ 的当前估计是 $\hat{g}_{[l-1]}(\cdot)$。我们使用 $\hat{g}_{[l]}(z)$ 来表示 $g(z)$ 的下一阶段估计量，其中 $\hat{g}_{[l]}(z)$ 和 $\hat{g}_{[l]}^{(1)}(z)$ 是如下方程的解（通过式(19.7)和(19.9)）： 580

$$\begin{aligned}0&=\sum_{i=1}^{N}\sum_{t=1}^{T}K_h(Z_{it},z)G_{it}(z,h)\\&\quad\times\mathcal{L}_{tg}(Y_i,\hat{g}_{[l-1]}(Z_{i1}),\cdots,\hat{g}_{[l]}(z)\\&\quad+[(Z_{it}-z)/h]'\hat{g}_{[l]}^{(1)}(z),\cdots,\hat{g}_{[l-1]}(Z_{iT}))\qquad(19.10)\\&=\sum_{i=1}^{N}\sum_{t=1}^{T}K_h(Z_{it},z)G_{it}(z,h)\Big[\sum_{s\neq t}\sigma^{ts}(Y_{is}-\hat{g}_{[l-1]}(Z_{is}))\\&\quad+\sigma^{tt}(Y_{it}-\hat{\alpha}_0(z)-[(Z_{it}-z)/h]'\hat{\alpha}_1(z))\Big]\qquad(19.11)\end{aligned}$$

方程(19.10)关于 $\hat{g}_{[l]}(z)$ 和 $\hat{g}_{[l]}^{(1)}(z)$ 是线性的，并且导致

$$\begin{bmatrix}\hat{g}_{[l]}(z)\\\hat{g}_{[l]}^{(1)}(z)\end{bmatrix}=D_1(z)^{-1}D_2(z)\tag{19.12}$$

其中

$$\begin{aligned}D_1(Z)&=\sum_{i=1}^{N}\sum_{t=1}^{T}K_h(Z_{it},z)\sigma^{tt}G_{it}(z,h)G_{it}(z,h)'\\D_2(Z)&=\sum_{i=1}^{N}\sum_{t=1}^{T}K_h(Z_{it},z)G_{it}(z,h)\\&\quad\times\Big[\sigma^{tt}Y_{it}+\sum_{s\neq t}^{T}\sigma^{ts}(Y_{is}-\hat{g}_{[l-1]}(Z_{is}))\Big]\end{aligned}$$

在 $g(\cdot)$ 估计的下一阶段，方程(19.12)得到 $\hat{g}_{[l]}(z)$，而 $\hat{g}_{[l]}^{(1)}(z)$ 是 $g^{(1)}(z)=\partial g(z)/\partial z$的下一步估计。直到收敛，我们获得 $g(z)$ 的最后估计 $\hat{g}(z)$。

上面的迭代过程要求 $g(\cdot)$ 的一个初始估计。可以选择标准局部常数或者忽略 Σ 的方差—协方差结构的局部线性估计。因为初始估计已经是 $g(\cdot)$ 的一个一致估计量，收敛通常在几步之后就会达到。

581 定义

$$\Omega(z) = -f(z)\sum_{t=1}^{T} E[\mathcal{L}_{i,ttg}(\cdot) \mid Z_{it} = z] = f(z)\sum_{t=1}^{T} E[\sigma^{tt} \mid Z_{it} = z] \tag{19.13}$$

其中第二个等式由式(19.8)得到。

同样，对 $r=1,\cdots,q$，定义 $b_r(z)$ 是由如下式子决定的一个有界平滑函数

$$b_r(z) = \frac{\kappa_2}{2}g_{rr}(z) + \Omega(z)^{-1}f(z)\sum_{t=1}^{T} E\left[\sum_{s\neq t}^{T}\sigma^{js}b_r(Z_{is}) \mid Z_{it} = z\right] \tag{19.14}$$

其中 $\kappa_2 = \int k(v)v^2\mathrm{d}v$, $g_{rr}(z) = \partial^2 g(z)/\partial z_r^2$。

在标准正则性条件下，例如 $g(\cdot)$ 是二次连续可微的，Wang(2003)证明

$$\hat{g}(z) - g(z) = \sum_{s=1}^{q} h_r^2 b_r(z) + v_N(z) + o_p\left(\sum_{s=1}^{q} h_r^2 + (Nh_1\cdots h_q)^{-1/2}\right) \tag{19.15}$$

其中

$$v_N(z) = \Omega(z)^{-1}\left\{\frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it},z)c_t'\Sigma^{-1}u_i\right\} \tag{19.16}$$

很显然，$E(v_N(Z))=0$，可以证明 $v_N(Z)$ 的渐近方差是(见习题19.2)

$$\mathrm{var}(v_N(z)) = \frac{\kappa^q}{\Omega(z)Nh_1\cdots h_q} + o(Nh_1\cdots h_q)^{-1} \tag{19.17}$$

其中 $\kappa = \int k^2(v)dv$。

假定 $h_\tau \propto N^{-1/(4+q)}$，利用式(19.14)和式(19.17)，可以证明

$$\sqrt{Nh_1\cdots h_q}\left(\hat{g}(z) - g(z) - \sum_{s=1}^{q} h_r^2 b_r(z)\right) \xrightarrow{d} N\left(0, \frac{\kappa^q}{\Omega(z)}\right) \tag{19.18}$$

其中 $\Omega(Z)$ 在式(19.13)中被定义。

Wang(2003)证明一次迭代(once-iterated)估计量与完全迭代(fully iterated)

582 估计量有相同的渐近行为。她的模拟也支持了这个结论，所以在实践中可以选择仅仅迭代一次。在这种情况下，偏误首项有一个闭合形式的表达式：

$$b_r(z) = \frac{1}{2}g_{rr}(z) + \Omega(z)^{-1}f(z)\sum_{t=1}^{T} E\left[\sum_{s\neq t}^{T}\sigma^{js}g_{rr}(Z_{is}) \mid Z_{it} = z\right] \tag{19.19}$$

接下来，我们考虑 Σ 的一个特定形式。我们假定 u_{it} 服从一个单向误差成分结构，即 $u_{it}=\mu_i+\nu_{it}$，其中 μ_i 是独立同分布 $(0,\sigma_\mu^2)$，ν_{it} 是独立同分布 $(0,\sigma_\nu^2)$。我们进一步假定 u_{it} 是条件同方差的。因此，$\Sigma=\text{cov}(u_i \mid Z_{i1},\cdots,Z_{iT})=\text{cov}(u_i)$ 给定为：

$$\Sigma = \sigma_\nu^2 I_T + \sigma_\mu^2 e_T e_T' = \sigma_\nu^2[I_T - \bar{J}_T] + \sigma_1^2 \bar{J}_T \tag{19.20}$$

其中 $\sigma_1^2 = T\sigma_\mu^2+\sigma_\nu^2$，$I_T$ 是一个 $T\times T$ 维的单位矩阵，e_T 是一个 $T\times 1$ 维的 1 的列向量，且 $\bar{J}_T = e_T e_T'/T$。很容易证实（参见 Wansbeek 和 Kapteyn(1982)）

$$\Sigma^{-1} = \frac{1}{\sigma_\nu^2}[I_T - \bar{J}_T] + \frac{1}{\sigma_1^2}\bar{J}_T$$

给定 $\mathcal{L}_i(\cdot)$ 关于 $g(\cdot)$ 是二次的，$\mathcal{L}_{i,tsg}$ 是一个常数：

$$\mathcal{L}_{i,ttg} = -c_t'\Sigma^{-1}c_t = -\sigma^{tt} \equiv -c_1$$

其中

$$c_1 \equiv \sigma^{tt} = \frac{1}{\sigma_\nu^2}\left(1-\frac{1}{T}\right) + \frac{1}{T\sigma_1^2} = \frac{(T-1)\sigma_\mu^2+\sigma_\nu^2}{\sigma_\nu^2\sigma_1^2}$$

因此，我们有

$$\Omega(z) = -f(z)\sum_{t=1}^{T} E[\mathcal{L}_{i,ttg}(\cdot) \mid Z_t = z] = c_1 Tf(z) \tag{19.21}$$

注意 $\mathcal{L}_{i,tsg}$ 同样是一个常数：

$$\mathcal{L}_{i,tsg} = -c_t'\Sigma^{-1}c_s = -\sigma^{ts} = -\frac{\sigma_\mu^2}{\sigma_1^2\sigma_\nu^2} \equiv -c_2,\quad t\neq s \tag{19.22}$$

583

因此，我们有

$$b_r(z) = \frac{\partial^2 g(z)}{\partial z_r^2} - \frac{c_2}{Tc_1}\sum_{t=1}^{T}\sum_{s\neq t}^{T} E[b_r(Z_{is}) \mid Z_{it} = z] \tag{19.23}$$

$(Nh_1\cdots h_q)^{1/2}\hat{g}(z)$ 的渐近方差被给出为

$$\begin{aligned}\text{avar}[(Nh_1\cdots h_q)^{1/2}\hat{g}(z)] &= \frac{\kappa^q}{c_1 Tf(z)} \\ &= \frac{\kappa^q\sigma_\nu^2\sigma_1^2}{Tf(z)[(T-1)\sigma_\mu^2+\sigma_\nu^2]} \\ &= \frac{\kappa^q\sigma_\nu^2}{Tf(z)}\left[1+\frac{\rho}{(T-1)\rho+1}\right]\end{aligned} \tag{19.24}$$

其中 $\rho = \frac{\sigma_\mu^2}{\sigma_\nu^2}$。

$g(z)$ 的传统核估计量忽略了 Σ 的方差结构，例如，由下式给定的局部常数

估计：

$$\bar{g}(z) = \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} Y_{it} K_h(Z_{it}, z)}{\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it}, z)} \tag{19.25}$$

有被下式给定的渐近方差[1]

$$\text{avar}[(Nh_1 \cdots h_q)^{1/2}\bar{g}(z)] = \frac{\kappa^q[\sigma_\nu^2 + \sigma_\mu^2]}{Tf(z)} = \frac{\kappa^q \sigma_\nu^2}{Tf(z)}[1+\rho] \tag{19.26}$$

相对渐近方差被给定为

$$\frac{\text{avar}[(Nh_1 \cdots h_q)^{1/2}\hat{g}(z)]}{\text{avar}[(Nh_1 \cdots h_q)^{1/2}\bar{g}(z)]} = \frac{1 + \dfrac{\rho}{(T-1)\rho+1}}{1+\rho} < 1 \tag{19.27}$$

其中 $\rho = \dfrac{\sigma_\mu^2}{\sigma_\nu^2}$。当 ρ 或者 T 增加时，以上比值变小。例如，对 $\rho = 1$，式(19.27)变为 $(T+1)/(2T)$，因为 $T \geqslant 2$，所以其值在 $[3/4, 1/2)$。对 $T = 2$ 且 ρ 为任意的情况，式(19.27)变为 $1 - \rho^2/(1+\rho)^2$，因为 $\rho \in (0, +\infty)$，所以其值在 $(0,1)$。对于较大的 ρ 值，使用 $\hat{g}(z)$ 替代 $\bar{g}(z)$ 得到的效率增进可能很大。

584

在实践中，σ_μ^2 和 σ_ν^2 是未知的，且需要被估计。可以证明 $\hat{\sigma}_\nu^2 = \dfrac{1}{N(T-1)} \sum_{i=1}^{N} \hat{u}_i'(I_T - \bar{J}_T)\hat{u}_i$ 是 σ_ν^2 的一个一致估计量且 $\hat{\sigma}_1^2 = \dfrac{1}{N}\sum_{i=1}^{N} \hat{u}_i' \bar{J}_T \hat{u}_i$ 是 σ_1^2 的一致估计，其中 $\hat{u}_i = (\hat{u}_{i1}, \cdots, \hat{u}_{iT})'$，$\hat{u}_{it} = Y_{it} - \bar{g}(Z_{it})$，$\bar{g}(Z_{it})$ 是在式(19.25)中给出的忽略 Σ 结构的 $g(Z_{it})$ 的初始估计。事实上，$\hat{\sigma}_\mu^2 - \sigma_\mu^2 = O_p(N^{-1/2})$，$\hat{\sigma}_1^2 - \sigma_1^2 = O_p(N^{-1/2})$。定义 $\hat{\Sigma}^{-1}$ 为 Σ^{-1}，把 σ_μ^2 和 σ_ν^2 替换为 $\hat{\sigma}_\mu^2$ 和 σ_ν^2，我们得到 $g(z)$ 的一个可行估计量。因为 $\hat{\Sigma}^{-1} - \Sigma^{-1} = O_p(N^{-1/2})$，所以，最终估计量的渐近分布与使用真正 Σ^{-1} 的估计量是一样的，因为 $\hat{\Sigma}^{-1} - \Sigma^{-1} = O_p(N^{-1/2})$，其收敛于零的速度快于非参数估计 $\hat{g}(z)$。

19.3 具有随机效应的部分线性模型

在本节，我们考虑一个具有随机效应的半参数部分线性面板数据回归模型：

---

[1] 局部线性估计量与局部常数核估计量有相同的渐近方差。具体参见第2章。

$$Y_{it} = X_{it}'\alpha + g(Z_{it}) + \mu_i + \nu_{it}, \quad i = 1,2,\cdots,N,\ t = 1,2,\cdots,T \tag{19.28}$$

其中 X_{it}是 $d\times 1$ 维的,而其他变量与在 19.2 节中定义的相同。

个体 i 的准则函数被修正为

$$\mathcal{L}_i(\cdot) = \mathcal{L}(Y_i, X_i, \alpha, g_i) = -\frac{1}{2}[Y_i - X_i'\alpha - g_i]'\Sigma^{-1}[Y_i - X_i'\alpha - g_i] \tag{19.29}$$

其中 $X_i = (X_{i1}, \cdots, X_{iT})'$。令$\mathcal{L}_{i,tg} = \partial\mathcal{L}_{i,tg}(\cdot)/\partial g_{it}$,那么我们有

$$\mathcal{L}_{i,tg} = -c_t'\Sigma^{-1}[Y_i - X_i'\alpha - g_i]$$

此时,假定 α 是已知的,则我们可以像 19.2 节那样简单地估计 $g(\cdot)$,但需将出现 Y_i 的地方换成 $Y_i - X_i'\alpha$。我们用 $\hat{g}(z,\alpha)$来表示相应的估计量。令 $\hat{g}_y(\cdot)$是式(19.12)的非参数估计量($\hat{g}_{[l]}(z)$),令 $\hat{g}_{x,r}(\cdot)$是式(19.12)中将 Y_{it}替换成 X_{it}第 r 个分量后的非参数估计量。更进一步地,令 $g_x(z) = \{g_{x,1}(z), \cdots, g_{x,d}(z)\}'$,有那么,立即很明显地通过平滑子(smoother)的线性化,并且由式(19.28)

$$\hat{g}(z,\alpha) = \hat{g}_y(z) - \hat{g}_x(z)'\alpha \tag{19.30}$$

这意味着

$$\frac{\partial}{\partial\alpha}\hat{g}(z,\alpha) = -\hat{g}_x(z)$$

定义 $\hat{g}_i(\alpha) = (\hat{g}(Z_{i1},\alpha), \cdots, \hat{g}(Z_{iT},\alpha))'$。对于 $\hat{g}_i(\alpha)$, $i = 1,2,\cdots,N$ 的给定估计值,我们通过最大化下式估计 α: 585

$$-\sum_{i=1}^{N}[Y_i - X_i'\alpha - \hat{g}_i(\alpha)]'\Sigma^{-1}[Y_i - X_i'\alpha - \hat{g}_i(\alpha)]$$

$$= -\sum_{i=1}^{N}[\hat{Y}_i - \hat{X}_i'\alpha]'\Sigma^{-1}[\hat{Y}_i - \hat{X}_i'\alpha] \tag{19.31}$$

其中 $\hat{Y}_i = (\hat{Y}_{i1}, \cdots, \hat{Y}_{iT})'$, $\hat{Y}_{it} = Y_{it} - \hat{g}_y(Z_{it})$,且 $\hat{X}_i = (\hat{X}_{i1}, \cdots, \hat{X}_{iT})'$, $\hat{X}_{it} = X_{it} - \hat{g}_x(Z_{it})$。

因为 $\hat{g}_y(Z_{it})$ 和 $\hat{g}_x(Z_{it})$ 都是与 α 无关的,我们能够使用 19.2 节讨论的方法来得到 $\hat{g}_y(Z_{it})$ 和 $\hat{g}_x(Z_{it})$ 。给定 $\hat{g}_y(Z_{it})$ 和 $\hat{g}_x(Z_{it})$,关于 α 最大化式(19.31)得到一个关于 $\hat{\alpha}$ 的闭式解:

$$\hat{\alpha} = \left[\sum_{i=1}^{N}\hat{X}_i\Sigma^{-1}\hat{X}_i\right]^{-1}\sum_{i=1}^{N}\hat{X}_i\Sigma^{-1}\hat{Y}_i \tag{19.32}$$

我们有

$$\sqrt{N}(\hat{\alpha} - \alpha) \to N(0,V) \text{ 依分布收敛}$$

其中 V 是一个正定矩阵。此外，V 的一个一致估计量为

$$\hat{V} = \left[N^{-1} \sum_{i=1}^{N} \hat{X}_i \Sigma^{-1} \hat{X}_i \right]^{-1}$$

Wang, Carroll 和 Lin(2005)证明，在 V 到达 α 的一个规则$\sqrt{n}$—致估计量的下界的意义上，$\hat{\alpha}$ 是 α 的一个半参数有效估计量。

给定 $\hat{\alpha}$，我们重写式(19.28)为

$$Y_{it} - X_{it}'\hat{\alpha} = g(Z_{it}) + \mu_i + \nu_{it} + X_{it}'(\alpha - \hat{\alpha}) \qquad (19.33)$$

方程(19.33)暗示我们能够使用在 19.2 节讨论过的方法来估计 $g(z)$，除了我们需要将 Y_{it} 替换成 $Y_{it} - X_{it}'\hat{\alpha}$。

更具体地，对于 $g(\cdot)$ 的当前阶段估计量($\hat{g}_{[l-1]}(\cdot)$)，我们通过下式估计 $g(\cdot)$ 的下一阶段估计量：

$$\begin{bmatrix} \hat{g}_{[l]}(z) \\ \hat{g}_{[l]}^{(1)}(z) \end{bmatrix} = D_1(z)^{-1} D_2(z, \hat{\alpha}) \qquad (19.34)$$

586 其中 $D_1(z)$ 与式(19.12)中定义的相同，即

$$D_1(z) = \sum_{i=1}^{N} \sum_{t=1}^{T} K_h(Z_{it}, z) \sigma^{tt} G_{it}(z,h) G_{it}(z,h)'$$

且 $D_2(z,\hat{\alpha})$ 被给定为

$$D_2(z,\hat{\alpha}) = \sum_{i=1}^{N} \sum_{t=1}^{T} K_h(Z_{it}, z) G_{it}(z,h) \times \left[\sigma^{tt} Y_{it} + \sum_{s \neq t}^{T} \sigma^{ts} (Y_{is} - X_{is}'\hat{\alpha} - \hat{g}_{[l-1]}(Z_{is})) \right]$$

直到收敛，我们得到 $\hat{g}(z)$ 的最终估计量(同时得到 $\hat{g}^{(1)}(z)$)。因为 $\hat{\alpha}$ 关于 α 是$\sqrt{N}$—致的，$\hat{g}(z)$ 的渐近分布与 19.2 节给出的是一样的(即与 α 已知的情况是相同的)。

19.4 具有固定效应的非参数面板数据模型

考虑如下具有固定效应的非参数面板数据回归模型：

$$Y_{it} = g(Z_{it}) + \mu_i + \nu_{it}, \quad i = 1,2,\cdots,N, \; t = 1,2,\cdots,T \qquad (19.35)$$

其中 $g(\cdot)$ 是一个未知的平滑函数，μ_i 是固定效应[①]，$Z_{it} = (Z_{it,1}, \cdots, Z_{it,q})$ 是 q 维的，所有其他变量是标量。考虑 N 较大、T 较小的情况，我们假定数据关于下

① 固定效应被认为是与具有未知相关结构的 Z_{it} 相关的随机误差项。

标 i 是独立的，并且我们假定，对所有 i、t 和 s，ν_{it} 独立于 X_{is}。

我们进行一阶差分来消除固定效应，即

$$Y_{it} - Y_{i1} = g(Z_{it}) - g(Z_{i1}) + \nu_{it} - \nu_{i1}, \quad i = 1,2,\cdots,N,\ t = 2,3,\cdots,T \tag{19.36}$$

我们考虑 ν_{it} 的两种情形：(i) ν_{it} 是具有零均值和有限(条件同)方差 σ_ν^2 的独立同分布；以及(ii) ν_{it} 具有零均值且可能存在序列相关，其相关结构未知。我们讨论 $g(\cdot)$ 的两种非参数估计，一种使用方差结构 Σ，另一种忽略 Σ 的结构。

587

19.4.1 误差方差结构已知

我们令 $\varepsilon_{it} = \nu_{it} - \nu_{i1}$，且令 $\varepsilon_i = (\varepsilon_{i1},\cdots,\varepsilon_{iT})'$。在本节我们假定误差 ν_{it} 是条件同方差，则 $\Sigma = \text{cov}(\varepsilon_i | Z_{i1},\cdots,Z_{iT}) = \text{cov}(\varepsilon_i)$，并且被给定为

$$\Sigma = \sigma_\nu^2(I_{T-1} + e_{T-1}e'_{T-1}) \tag{19.37}$$

其中 I_{T-1} 是 $(T-1)\times(T-1)$ 维的单位矩阵，e_{T-1} 是一个 $(T-1)\times 1$ 维的1向量。很容易验证

$$\Sigma^{-1} = \frac{1}{\sigma_\nu^2}\left[I_{T-1} - \frac{1}{T}e_{T-1}e'_{T-1}\right] \tag{19.38}$$

Henderson，Carroll 和 Li(2006)建议通过为个体 i 选择准则函数而使用一种剖面似然方法：

$$\mathcal{L}_i(\cdot) = \mathcal{L}(Y_i, g_i) = -\frac{1}{2}[\tilde{Y}_i - g_i + g_{i1}e_{T-1}]'\Sigma^{-1}[\tilde{Y}_i - g_i + g_{i1}e_{T-1}] \tag{19.39}$$

其中 $\tilde{Y}_i = (\tilde{Y}_{i2},\cdots,\tilde{Y}_{iT})'$，$\tilde{Y}_{it} = Y_{it} - Y_{i1}$，$g_{it} = g(Z_{it})$，且 $g_i = (g_{i2},\cdots,g_{iT})'$。

定义乘积核

$$K_h(v) = \prod_{j=1}^{q} h_j^{-1}k(v_j/h_j)$$

其中 $k(\cdot)$ 是一个一元核函数。令 $f(\cdot)$ 表示 Z_{it} 的概率密度函数，令 $(Z_{it}-z)/h) = [(Z_{it,1}-Z_1)/h_1,\cdots,(Z_{it,q}-Z_q)/h_q)]'$ 且定义 $G_{it}(z,h) = (1,[(Z_{it}-z)/h]')'$ (G_{it} 是 $(q+1)\times 1$ 维)，令 $g^{(1)}(z) = \partial g(z)/\partial z$ 是 $g(\cdot)$ 关于 Z 的一阶偏导数($q+1$ 维)。我们通过 $\alpha_0(z)$ 来估计未知函数 $g(z)$，其中 $\alpha_0(z)$ 和 $\alpha_1(z)$ 求解了下面的一阶条件：

$$\begin{aligned} 0 = \sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it},z)G_{it}(z,h) \times \mathcal{L}_{i,tg}(Y_i, \hat{g}(Z_{i1}),\cdots,\alpha_0(z) \\ + [(Z_{it}-z)/h]'\alpha_1(z),\cdots,\hat{g}(Z_{iT})) \end{aligned} \tag{19.40}$$

其中$\mathcal{L}_{i,tg}=\partial\mathcal{L}_i(\cdot)/\partial g_{it}$。当$s\neq t$时，$\mathcal{L}_{i,tg}$中的变量是$g(Z_{is})$，而当$s=t$时，则为$\alpha_0(z)+[(Z_{it}-z)/h]'\alpha_1(z)$。

注意式(19.40)是$(q+1)\times 1$维的，从中我们求解得$(q+1)$个参数$\hat{\alpha}_0(z)$（一个标量）及$\hat{\alpha}_1(z)$（一个$q\times 1$向量）。

对非参数估计的一个迭代过程

588 Henderson 等(2006)证明可以使用如下的迭代算法来估计$g(\cdot)$，其中我们也需要施加约束$\sum_{i=1}^{N}\sum_{t=1}^{T}\{Y_{it}-\hat{g}(Z_{it})\}=0$来保证基于式(19.36)的$g(\cdot)$是唯一被定义的。算法关于$Y_{it}$是线性的且被描述如下。

定义一些当前阶段的变量（在$l-1$阶段）为

$$H_{i,[l-1]}=\begin{bmatrix} Y_{i2}-\hat{g}_{[l-1]}(Z_{i2})\\ \vdots \\ Y_{iT}-\hat{g}_{[l-1]}(Z_{iT})\end{bmatrix}-[Y_{i1}-\hat{g}_{[l-1]}(Z_{i1})]e_{T-1}$$

$$D_1=\frac{1}{N}\sum_{i=1}^{N}\Big[e'_{T-1}\Sigma^{-1}e_{T-1}K_h(Z_{i1},z)G_{i1}G'_{i1}+\sum_{t=2}^{T}c'_{t-1}\Sigma^{-1}c_{t-1}K_h(Z_{it},z)G_{it}G'_{it}\Big]$$

$$D_{2,[l-1]}=\frac{1}{N}\sum_{i=1}^{N}\Big[e'_{T-1}\Sigma^{-1}e_{T-1}K_h(Z_{i1},z)G_{i1}\hat{g}_{[l-1]}(Z_{i1})+\sum_{t=2}^{T}c'_{t-1}\Sigma^{-1}c_{t-1}K_h(Z_{it},z)G_{it}\hat{g}_{[l-1]}(Z_{it})\Big]$$

$$D_{3,[l-1]}=\frac{1}{N}\sum_{t=1}^{N}\Big[\sum_{t=2}^{T}K_h(Z_{it},z)G_{it}c'_{t-1}\Sigma^{-1}H_{i,[l-1]}-K_h(Z_{i1},z)G_{i1}e'_{T-1}\Sigma^{-1}H_{i,[l-1]}\Big] \tag{19.41}$$

Henderson 等(2006)证明下一个阶段的估计量（l阶段）为（见习题 19.3）

$$\begin{bmatrix}\hat{g}_{[l]}(z)\\ \hat{g}^{(1)}_{[l]}(z)\end{bmatrix}=D_1^{-1}[D_{2,[l-1]}+D_{3,[l-1]}] \tag{19.42}$$

在实践中，Σ是未知的，在初始估计中可以通过一个单位矩阵来替换Σ。然而，即使Σ被替换成一个单位矩阵，式(19.42)仍然是一个迭代估计过程。Henderson 等(2006)证明这个过程经常在三到四步后收敛。此外，可以使用非参数

589 序列分析法来获得$g(\cdot)$的一个一致初始估计。我们将在 19.8 节讨论固定效应模型的非参数序列估计。

令 $b_r(z)$ 是一个有界和连续的函数,它是下式的解

$$b_r(z) = \frac{\kappa_2}{2} g_{rr}(z) - \frac{\sigma_\nu^2 f(z)}{T-1} \sum_{t=1}^{T} \sum_{s \neq t}^{T} E[\mathcal{L}_{i,tsg}(\cdot) b_r(Z_{is}) \mid Z_{it} = z] \tag{19.43}$$

其中 $g_{rr}(z) = \partial^2 g(z)/\partial z_r^2 (r = 1, \cdots, q)$,且 $\mathcal{L}_{i,tsg} = \partial^2 \mathcal{L}_i(\cdot)/(\partial g_{it} \partial g_{is})$。很容易证明(见习题 19.4)

$$\mathcal{L}_{i,1tg} = -c_{t-1}' \Sigma^{-1} e_{T-1} = -\frac{1}{T\sigma_\nu^2}, t \geq 2 \tag{19.44}$$

且

$$\mathcal{L}_{i,tsg} = -c_{t-1}' \Sigma^{-1} c_{s-1} = \frac{1}{T\sigma_\nu^2}, t, s \geq 2, t \neq s \tag{19.45}$$

一般而言,$b_r(\cdot)$ 没有一个封闭形式的表达式。但是,若使用 $g(\cdot)$ 的一个一致初始估计量,则对于 $b_r(\cdot)$ 的一个一步迭代有一个封闭形式的表达式,且被给定为(在方程的右边把 $b_r(\cdot)$ 替换成 $g_{rr}(\cdot)$)

$$b_r(z) = \frac{\kappa_2}{2} g_{rr}(z) - \frac{\sigma_\nu^2 f(z)}{T-1} \sum_{t=1}^{T} \sum_{s \neq t}^{T} E[\mathcal{L}_{i,tsg}(\cdot) g_{rr}(Z_{is}) \mid Z_{it} = z] \tag{19.46}$$

令 $\hat{g}(z)$ 表示 $g(z)$ 的收敛估计量, Henderson 等(2006)证明

$$E[\hat{g}(z)] - g(z) = (\kappa_2/2) \sum_{r=1}^{q} h_r^2 b_r(z) + o\left(\sum_{r=1}^{q} h_r^2\right)$$

$$\operatorname{var}(\hat{g}(z)) = \frac{\kappa^q}{Nh_1 \cdots h_q} \frac{\sigma_\nu^2}{(T-1)f(z)} + o((Nh_1 \cdots h_q)^{-1}) \tag{19.47}$$

在 $h_r \propto N^{-1/(4+q)}$ 的假定下,我们有如下关于 $\hat{g}(z)$ 的渐近分布($\kappa = \int k^2(v)\mathrm{d}v$):

$$\sqrt{Nh_1 \cdots h_q}\left[\hat{g}(z) - g(z) - \sum_{r=1}^{q} h_r^2 b_r(z)\right] \xrightarrow{d} N\left(0, \frac{\sigma_\nu^2}{(T-1)f(z)}\right)$$

590

很显然,$f(z)$ 和 σ_ν^2 可以通过下式来一致估计:

$$\hat{f}(z) = (NT)^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} K_h(Z_{it}, z)$$

$$\hat{\sigma}_\nu^2 = \frac{1}{2N(T-1)} \sum_{i=1}^{N} \sum_{t=2}^{T} \hat{u}_{it}^2$$

$$\hat{u}_{it} = Y_{it} - Y_{i1} - (\hat{g}(Z_{it}) - \hat{g}(Z_{i1}))$$

为了估计偏误首项,我们需要 $b_r(z)$ 的一个一致估计量:

$$\hat{b}_r(z) = \frac{\kappa_2}{2} \hat{g}_{rr}(z) - \frac{\hat{\sigma}_\nu^2}{N(T-1)} \sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{s \neq t}^{T} K_h(Z_{it}, z) \mathcal{L}_{i,tsg} \hat{g}_{rr}(Z_{is})$$

其中 $\hat{g}_{rr}(z)$ 是 $g_{rr}(z)$ 的一个一致估计量，其可以通过比如基于式(19.36)的局部二次回归来获得。

19.4.2 误差方差结构未知

当 Σ 是未知形式时，仍然可以通过估计每一个成分 $\sigma_{jl}, 1 \leqslant j \leqslant l \leqslant T$ 来获得 Σ 的一个一致估计。因为 T 是有限的，仅仅只有有限多个参数需要被估计；因此，得到的估计量将是 $\sqrt{N}$ 一致的。当我们使用 Σ^{-1} 的一个 $\sqrt{N}$ 一致估计量时，在最后一节描述的估计量 $\hat{g}(z)$ 的渐近分布仍然没有改变。但是，人们也可以选择忽略 Σ 的方差结构。在本节，我们忽略 Σ 的方差结构来推导一个固定效应估计量的渐近分布。在这种情况下，式(19.39)的目标函数被修正如下(将 Σ^{-1} 替换为 I_{T-1})

$$\mathcal{L}_i(\cdot) = \mathcal{L}(Y_i, g_i) = -\frac{1}{2}[\tilde{Y}_i - g_i + g_{i1}e_{T-1}]'[\tilde{Y}_i - g_i + g_{i1}e_{T-1}]' \tag{19.48}$$

如式(19.41)给出的 D_1、D_2 和 D_3 的定义，但现在将 Σ^{-1} 替换为 I_{T-1}，我们定义(其中我们也使用 $e'_{T-1}e_{T-1} = T-1$ 及 $c'_{t-1}c_{t-1} = 1$)

591

$$J_1 = \frac{T-1}{N}\sum_{i=1}^{N}\left[K_h(Z_{i1}, z)G_{i1}G'_{i1} + \sum_{t=2}^{T}K_h(Z_{it}, z)G_{it}G'_{it}\right]$$

$$J_{2,[l-1]} = \frac{T-1}{N}\sum_{i=1}^{N}\left[K_h(Z_{i1}, z)G_{i1}\hat{g}_{[l-1]}(Z_{i1}) + \sum_{t=2}^{T}K_h(Z_{it}, z)G_{it}\hat{g}_{[l-1]}(Z_{it})\right]$$

$$J_{3,[l-1]} = \frac{1}{N}\sum_{i=1}^{N}\left[\sum_{t=2}^{T}K_h(Z_{it}, z)G_{it}c'_{t-1}H_{i,[l-1]} - K_h(Z_{i1}, z)G_{i1}e'_{T-1}H_{i,[l-1]}\right]$$

$g(z)$ 和 $g^{(1)}(z)$ 的下一步估计量为

$$\begin{bmatrix} \tilde{g}_{[l]}(z) \\ \tilde{g}^{(1)}_{[l]}(z) \end{bmatrix} = J_1^{-1}[J_{2,[l-1]} + J_{3,[l-1]}] \tag{19.49}$$

可以使用 19.4.1 节给出的结论通过将 Σ^{-1} 替换为 I_{T-1} 来推导出 $\tilde{g}(z)$ 的渐近分布。特别是，偏误首项仍然由式(19.43)给出，且对所有 t, s 有 $L_{i,ts} = -1$(通过在式(19.44)和式(19.45)中将 Σ^{-1} 替换为 I_{T-1})。Henderson 等(2006)证明了 $\tilde{g}(z)$ 的渐近方差被给定为

$$\mathrm{var}(\tilde{g}(z)) = \frac{\sigma_\nu^2 \kappa^q (2+T)}{4(T-1)f(z)Nh_1\cdots h_q} + o((Nh_1\cdots h_q)^{-1}) \quad (19.50)$$

Henderson 等(2006)同样提出方法来检验一个非参数随机效应模型与一个非参数固定效应模型(即一个“非参数 Hausman 检验”)。令 $\hat{g}_{ran}(\cdot)$ 和 $\hat{g}_{fix}(\cdot)$ 表示 $g(\cdot)$ 非参数的随机和固定效应估计，Henderson 等提出的检验统计量是基于 $\hat{g}_{ran}(\cdot)$ 和 $\hat{g}_{fix}(\cdot)$ 之间的均方差：

$$\frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T}[\hat{g}_{ran}(Z_{it}) - \hat{g}_{fix}(Z_{it})]^2$$

通过使用一个基于残差的自举法可以获得一个自举对应物(原始自举法)。

19.5　一个固定效应部分线性模型

如第 7 章所提到的,非参数回归遭遇了维数诅咒,因此,当面临大量回归元时,实践者经常倾向于半参数方法。在本节,我们考虑一个半参数部分线性模型,只有回归元的一个子集非参数地进入模型。一个具有固定效应的部分线性面板数据回归模型被给定为　592

$$Y_{it} = X_{it}'\alpha + g(Z_{it}) + \mu_i + \nu_{it}, \quad i = 1,2,\cdots,N,\ t = 1,2,\cdots,T \quad (19.51)$$

其中 X_{it} 和 α 是 $d\times 1$ 维的,其他剩余变量与在 19.4 节定义的相同。

我们进行一阶差分来消除固定效应,于是

$$\tilde{Y}_{it} = \tilde{X}_{it}'\alpha + g(Z_{it}) - g(Z_{i1}) + \varepsilon_{it}, \quad i = 1,2,\cdots,N,\ t = 2,3,\cdots,T \quad (19.52)$$

其中 $\tilde{Y}_{it} \equiv \tilde{Y}_{it} - Y_{i1}$，$\tilde{X}_{it} \equiv X_{it} - X_{i1}$，$\varepsilon_{it} = \nu_{it} - \nu_{i1}$。

个体 i 的准则函数被修正如下：

$$\begin{aligned}\mathcal{L}_i(\cdot) &= \mathcal{L}(Y_i, X_i, \alpha, g_i)\\ &= -\frac{1}{2}[\tilde{Y}_i - \tilde{X}_i'\alpha - g_i + g_{i1}e_{T-1}]'\Sigma^{-1}[\tilde{Y}_i - \tilde{X}_i'\alpha - g_i + g_{i1}e_{T-1}]\end{aligned} \quad (19.53)$$

其中 $\tilde{X}_i = (\tilde{X}_{i2},\cdots,\tilde{X}_{iT})'$。导函数变为

$$\mathcal{L}_{i,1g} = -e_{T-1}\Sigma^{-1}[\tilde{Y}_{i,-1} - \tilde{X}_i\alpha - g_i + g_{i1}e_{T-1}]$$

且

$$\mathcal{L}_{i,tg} = c_{t-1}\Sigma^{-1}[\tilde{Y}_i - \tilde{X}_i\alpha - g_i - g_{i1}e_{T-1}], \quad t \geqslant 2$$

Henderson 等(2006)建议通过一个剖面核方法来估计 $g(\cdot)$ 和 α。对于 α 的一个给定值以及当前阶段估计量 $\hat{g}_{[l-1]}(\cdot)$,我们通过 $\hat{g}_{[l]}(z)$ 来估计下一步的 $g(z)$,其中 $\hat{g}_{[l]}(z)$ 以及 $\hat{g}_{[l]}^{(1)}(z)$ 满足一阶条件

593

$$0 = \sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it},z)G_{it}(z,h)\,\mathcal{L}_{tg}(Y_i,X_i,\alpha,\hat{g}_{[l-1]}(Z_{i1},\alpha),\cdots,\hat{g}_{[l]}(z) + [(Z_{it}-z)/h]'\hat{g}_{[l]}^{(1)}(z),\cdots,\hat{g}(Z_{iT,\alpha})) \qquad (19.54)$$

将式(19.52)与式(19.36)进行比较,令 $\hat{g}_y(\cdot)$ 是式(19.36)中的非参数估计量,令 $\hat{g}_{x,r}(\cdot)$ 是式(19.36)中将 Y_{it} 换成 X_{it} 的第 r 个成分后的非参数估计量。更进一步地,令 $\hat{g}_x(z)=\{\hat{g}_{x,1}(z),\cdots,\hat{g}_{x,d}(z)\}'$,那么;通过平滑子的线性化,并且由式(19.52),明显地有

$$\hat{g}(z,\alpha) = \hat{g}_y(z) - \hat{g}_x(z)'\alpha \qquad (19.55)$$

定义 $\hat{Y}_{it}^{*} = \tilde{Y}_{it} - \{\hat{g}_y(Z_{it}) - \hat{g}_y(Z_{i1})\}$, $\hat{Y}_i^{*} = (\hat{Y}_{i2}^{*},\cdots,\hat{Y}_{iT}^{*})'$, $\hat{X}_{it}^{*} = \tilde{X}_{it} - \{\hat{g}_x(Z_{it}) - \hat{g}_x(Z_{i1})\}$ 以及 $\hat{X}_{it}^{*} = (\hat{X}_{i2}^{*},\cdots,\hat{X}_{iT}^{*})'$。Henderson 等(2006)建议通过最小化下式来估计 α:

$$\sum_{i=1}^{N}[\hat{Y}_i^{*} - \hat{X}_i^{*\prime}\alpha]'\Sigma^{-1}[\hat{Y}_i^{*} - \hat{X}_i^{*\prime}\alpha] \qquad (19.56)$$

从而导致如下的闭式解:

$$\hat{\alpha} = \left[\sum_{i=1}^{N}\hat{X}_i^{*}\Sigma^{-1}\hat{X}_i^{*\prime}\right]^{-1}\sum_{i=1}^{N}\hat{X}_i^{*}\Sigma^{-1}\hat{Y}_i^{*} \qquad (19.57)$$

注意因为我们有对于 $\hat{\alpha}$ 的一个闭式解,没必要用迭代来估计 $\hat{\alpha}$。

$\hat{\alpha}$ 的渐近分布是

$$\sqrt{N}(\hat{\alpha}-\alpha) \to N(0,V_1) \text{ 依分布收敛} \qquad (19.58)$$

其中 V_1 是一个正定矩阵。此外,V_1 的一个一致估计量被给定为

$$\hat{V}_1 = \left[N^{-1}\sum_{i=1}^{N}\hat{X}_i^{*}\Sigma^{-1}\hat{X}_i^{*\prime}\right]^{-1}$$

$\hat{\alpha}$ 由式(19.57)给定,我们可以用 19.4 节中介绍的方法基于下式来估计 $g(\cdot)$,除了现在我们需要将出现 $\tilde{Y}_{it}$ 的地方换成 $\tilde{Y}_{it} - \tilde{X}_{it}'\hat{\alpha}$:

$$\tilde{Y}_{it} - \tilde{X}_{it}'\hat{\alpha} = g(Z_{it}) - g(Z_{i1}) + \text{error} \qquad (19.59)$$

594

直到收敛,得到的 $\hat{g}(z)$ 具有与 19.4 节中给出的相同的渐近分布。这是因为通过用真实的 α 来替换式(19.59)中的 $\hat{\alpha}$ 简便地给出了 19.4 节所涵盖的非参数回归模型。接下来,注意 $\hat{\alpha}-\alpha = O_p(N^{-1/2})$,其收敛于零的速度快于非参数估计量 $\hat{g}(z)-g(z)$。因此,通过用 $\hat{\alpha}$ 来替换真实的 α 将不会影响 $\hat{g}(z)$ 的渐近分布。

也可以通过忽略方差结构 Σ 来估计 α。这种方法也有计算简便的优势。令 $\tilde{Y}_i^*$ 和 $\tilde{X}_i^*$ 表示 $N\times 1$ 和 $N\times d$ 矩阵，其第 i 行分别被给定为 $\tilde{Y}_{it}^* = \tilde{Y}_{it} - [\tilde{g}_y(Y_{it}) - \tilde{g}_y(Z_{i1})]$ 和 $\tilde{X}_{it}^* = \tilde{X}_{it} - [\tilde{g}_x(Z_{it}) - \tilde{g}_x(Z_{i1})]$，其中 $g_y(\cdot)$ 和 $\tilde{g}_x(\cdot)$ 以与 $\hat{g}_y(\cdot)$ 和 $\hat{g}_x(\cdot)$ 相同的方式定义，除了我们将 Σ^{-1} 替换为 I_{T-1}。那么，可以通过下式估计 α

$$\hat{\alpha} = \left[\sum_{i=1}^{N}\tilde{X}_i^*\tilde{X}_i^{*\prime}\right]^{-1}\left[\sum_{i=1}^{N}\tilde{X}_i^*\tilde{Y}_i^*\right] \tag{19.60}$$

$\tilde{\alpha}$ 的渐近分布被给定为

$$\sqrt{N}(\tilde{\alpha}-\alpha)\rightarrow N(0,V_2)\text{ 依分布收敛} \tag{19.61}$$

其中 $V_2 = A^{-1}BA^{-1}$ 是一个正定矩阵。此外，$\hat{V}_2 = \hat{A}^{-1}\hat{B}\hat{A}^{-1}$ 是 V_2 的一个一致估计量，其中

$$\hat{A} = N^{-1}\sum_{i=1}^{N}\tilde{x}_i^*\tilde{x}_i^{*\prime}$$

$$\hat{B} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}\tilde{\varepsilon}_{it}\tilde{\varepsilon}_{is}\tilde{X}_{it}^*\tilde{X}_{is}^*$$

其中 $\tilde{\varepsilon}_{it} = \tilde{Y}_{it}^* - \tilde{X}_{it}^{*\prime}\tilde{\alpha}$。

19.6　半参数工具变量估计量

在这一节我们讨论形如式(19.1)的一个部分线性模型的估计，其中 X_{it} 可能与误差项 u_{it} 相关。我们允许误差项是序列相关的(比如，一个随机效应模型)，但在这一节我们排除固定效应模型。

19.6.1　一个不可行的估计量

到目前为止，对于具有一个单向随机效应误差结构的部分线性面板数据模型(19.1)，我们仅仅考虑了 X_{it} 和 Z_{it} 都是严格外生的情况。在本节，我们允许 X_{it} 是内生的可能性。然而，我们仍然假定 Z_{it} 是严格外生的。对于允许 Z_{it} 是内生的一个部分线性模型，参见第 16 章的 16.3.3 节。像在第 7 章那样，我们首先消除 $g(\cdot)$。对式(19.1)关于 Z_{it} 取条件期望，然后从式(19.1)中减去它，得到　595

$$Y_{it} - E(Y_{it}\mid Z_{it}) = (X_{it} - E(X_{it}\mid Z_{it}))'\alpha + u_{it} \equiv v_{it}'\alpha + u_{it} \tag{19.62}$$

其中我们使用了

$$E(u_{it}\mid Z_{it}) = 0, v_{it} \stackrel{\text{def}}{=} X_{it} - E(X_{it}\mid Z_{it})$$

我们允许 X_{it}（因此 v_{it}）与 u_{it} 相关的可能性。若 v_{it} 包含 Y_{it} 的滞后值和/或 u_{it} 是序列相关将出现这种情况。假定存在一个与 X_{it} 相关且与 u_{it} 不相关的 $p\times 1$（$p\geq q, X_{it}\in\mathbb{R}^q$）工具向量 W_{it}（详见下面的式(19.64)）。那么，我们能够使用工具变量最小二乘估计 α 得到

$$\tilde{\alpha}_{\text{IV}}=(v'ww'v)^{-1}v'ww'(y-E(y|z))=\alpha+(v'ww'v)^{-1}v'ww'u \quad (19.63)$$

其中 w 和 v 是 $N\times p$ 和 $N\times q$ 维的，具有典型行 W_{it}' 和 v_{it}'。同样，y、$E(y|z)$ 和 u 都是 $N\times 1$ 向量，有典型行元素给定为 Y_{it}、$E(Y_{it}|Z_{it})$ 和 u_{it}。令 W_i 和 u_i 表示 $T\times p$ 和 $T\times 1$ 矩阵，其典型行由 W_{it}' 和 u_{it} 给出。我们需要如下的条件来推导 $\tilde{\alpha}_{\text{IV}}(n=NT)$ 的渐近分布：

$$w'v/n\xrightarrow{p}E[W_{it}v_{it}']\equiv A$$

$$(w'uu'w)/n\xrightarrow{p}E[W_i'u_iu_i'W_i]/T\equiv B, B\text{ 是正定的}$$

$$w'u/n\xrightarrow{p}E[W_{it}u_{it}]=0 \quad (19.64)$$

使用式(19.64)和中心极限定理，很容易证明

$$\sqrt{N}(\tilde{\alpha}_{IV}-\alpha)\rightarrow N(0,V)\text{ 依分布收敛} \quad (19.65)$$

其中 $V=Q^{-1}A'BAQ^{-1}$，$Q=A'A$，A 和 B 在式(19.64)被定义。

当 X_{it} 实际上是弱外生的，那么可以简单地选择 $W_{it}=X_{it}-E(X_{it}|Z_{it})=v_{it}$，在式(19.63)中定义的 $\tilde{\alpha}$ 变成第7章讨论的式(7.5)中定义的（不可行）估计量。

19.6.2 工具变量的选择

我们现在转向如何选择工具 W_{it} 这个基本的问题。考虑 X_{it} 和 Z_{it} 都是标量这种最简单的情况，其中 $X_{it}=Y_{i,t-1}$ 和 Z_{it} 是外生的。注意我们一定不能选择 $W_{it}=Z_{i,t-1}$ 作为 $v_{it}=Y_{i,t-1}-E(Y_{i,t-1}|Z_{it})$ 的一个工具变量。因为即使 $Z_{i,t-1}$ 与 u_{it} 是不相关的，由于 $g(\cdot)$ 的函数形式是未知的，所以不能保证 $Z_{i,t-1}$ 与 v_{it} 是相关的，因此 $Z_{i,t-1}$ 与 v_{it} 可能是弱相关甚至不相关的。通过 Z_{it} 是外生的假定，我们知道 $E(u_{it}|Z_{i,t-1})=0$，所以 $Z_{i,t-1}$ 与误差项 u_{it} 是不相关的。但是我们同样能够有 $E(v_{it}Z_{i,t-1})=0$，以致 $Z_{i,t-1}$ 与 v_{it} 是不相关的。因此，$Z_{i,t-1}$ 不是一个合理的工具变量。为了表明这确实是可能的，考虑 Z_{it} 是一个独立的分布过程这种情形，满足

596

$$E(Z_{it})=0,\quad E(Z_{it}^3)=0 \quad (19.66)$$

同样假定 $g(Z_{it})=Z_{it}^2$。在这种情况下，我们有

$$Y_{it} = \alpha Y_{i,t-1} + Z_{it}^2 + u_{it} \tag{19.67}$$

从式(19.67)中很容易发现 $Y_{i,t-1}$ 与 Z_{it} 是彼此独立的,因此

$$E(Y_{i,t-1} \mid Z_{it}) = E(Y_{i,t-1}) \stackrel{\text{def}}{=} \mu_y$$

利用式(19.66)、(19.67)以及 $E(u_{i,t-1} \mid Z_{i,t-1}) = 0$,我们有

$$E[v_{it}Z_{i,t-1}] = \alpha E(Y_{i,t-2})E(Z_{i,t-1}) + E(Z_{i,t-1}^3) + E(u_{i,t-1}Z_{i,t-1}) = 0 \tag{19.68}$$

这表明 $Z_{i,t-1}$ 与内生回归量 v_{it} 是不相关的,因此不可能是一个合理的工具变量。当然为了表明 $Z_{i,t-1}$ 不是一个合理的工具变量,我们已经设定了一些较强的假定。在实践中,$Z_{i,t-1}$ 与 v_{it} 不相关是不可能的,但 $Z_{i,t-1}$ 与 v_{it} 之间弱相关肯定是可能的。

给定式(19.62)的结构及 $X_{it} = Y_{i,t-1}$,我们知道 $v_{it} = Y_{i,t-1} - E(Y_{i,t-1} \mid Z_{it})$ 一定与 $Z_{i,t-1}$ 的一些函数相关,虽然它可能不与 $Z_{i,t-1}$ 的特定线性函数相关。Newey(1900a)讨论了在独立数据的一般(参数)非线性回归模型中的最优工具变量估计。若我们考虑不现实的情况,如 $E(Y_{it} \mid Z_{it})$ 和 $E(X_{it} \mid Z_{it})$ 是已知的,那么 $v_{it} = X_{it} - E(X_{it} \mid Z_{it})$ 就是已知的。在这种情况下,式(19.62)是一个线性参数模型,并且我们可以使用 Newey 的结论来获得一个与内生变量 v_{it} 最大相关的工具变量。若我们限制自己必须选择的工具是 $Z_{i,t-1}$ 的函数,那么,适合的工具简单地就是 v_{it} 在 $Z_{i,t-1}$ 上的最优投射,或者是给定 $Z_{i,t-1}$ 下 v_{it} 的条件均值函数。于是,被提出的工具是

597

$$\begin{aligned} W_{it} &= E(v_{it} \mid Z_{i,t-1}) \\ &= E(Y_{i,t-1} \mid Z_{i,t-1}) - E[E(Y_{i,t-1} \mid Z_{it}) \mid Z_{i,t-1}] \end{aligned} \tag{19.69}$$

在实践中,这些条件期望是未知的,并且我们必须使用非参数方法来估计它们。方程(19.69)是有计算要求的,因为它包含了一个双重的条件期望,即 $E[E(\cdot \mid Z_{it}) \mid Z_{i,t-1}]$。然而,注意

$$\begin{aligned} Y_{i,t-1} &= [Y_{i,t-1} - E(Y_{i,t-1} \mid Z_{it})] + E(Y_{i,t-1} \mid Z_{it}) \\ &\equiv v_{it} + E(Y_{i,t-1} \mid Z_{it}) \end{aligned}$$

且 $E(Y_{i,t-1} \mid Z_{it})$ 与 v_{it} 是正交的,因此,

$$\begin{aligned} E[v_{it}E(Y_{i,t-1} \mid Z_{i,t-1})] &= E(v_{it}E[v_{it} + E(Y_{i,t-1} \mid Z_{it}) \mid Z_{i,t-1}]) \\ &= E(v_{it}[E(v_{it} \mid Z_{i,t-1}) + \mu_y]) \\ &= E(v_{it}E[v_{it} \mid Z_{i,t-1}]) + \mu_y E[v_{it}] \\ &= E(v_{it}E[v_{it} \mid Z_{i,t-1}]) \end{aligned} \tag{19.70}$$

其中我们使用了如下的假定:Z_{it} 是一个独立同分布过程,有

$$E(Y_{i,t-1} \mid Z_{it}) = E(Y_{i,t-1}) = \mu_y$$

以及$E(v_{it})=0$的事实。方程(19.70)告诉我们，$E(Y_{i,t-1}|Z_{i,t-1})$相当于$E(v_{it}|Z_{i,t-1})$的一个等价工具变量。但是，$E(Y_{i,t-1}|Z_{i,t-1})$是基于一个单一的条件期望，所以比包含一个双重条件期望（见式(19.70)）的$E(v_{it}|Z_{i,t-1})$更容易计算。因此，我们建议使用

$$W_{it} = E(Y_{i,t-1} \mid Z_{i,t-1})$$

作为v_{it}的一个工具变量而非使用式(19.70)。

甚至当Z_{it}不是一个独立同分布过程时，比如当(Y_{it},Z_{it})关于t是平稳的，也很容易证明$E(Y_{i,t-1}|Z_{i,t-1})$和v_{it}是正相关的，因此$E(Y_{i,t-1}|Z_{i,t-1})$是一个合理的工具变量。

在一般情形下，我们有$X_{it}=(Y_{i,t-1},X'_{2,it})'$，其中$X_{2,it}$是外生的。那么，我们的工具变量将是$W_{it}=(W_{1,it},W'_{2,it})'$，其中

$$W_{1,it} = E(Y_{i,t-1} \mid Z_{i,t-1}), \quad W_{2,it} = X_{2,it} - E(X_{2,it} \mid Z_{it})$$

（其等价于$X_{2,it}$）。

19.6.3 一个可行的估计量

19.6.1节介绍的估计量$\tilde{\alpha}_{IV}$因为条件均值函数$E(Y_{it}|Z_{it})$、$E(X_{it}|Z_{it})$以及

598 $E(W_{it}|Z_{it})$是未知的，所以不可行。通过将未知的条件均值函数替换为非参数估计量，能够获得一个可行的估计量，比如核估计。接下来，我们将使用核方法来估计如上所示的未知条件期望。具体地，我们通过$\hat{f}_{it}$、$\hat{\xi}_{it}$、$\hat{X}_{it}$以及$\hat{W}_{it}$来估计$f(Z_{it})$，$\xi_{it}\equiv E(Y_{i,t}|Z_{it})$，$E(X_{it}|Z_{it})$，$W_{it}=E(X_{it}|Z_{i,t-1})$，其中（回忆$Z_{it}\in\mathbb{R}^q$）

$$\hat{f}_{it} = \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} K_{h,it,js}$$

$$\hat{\xi}_{it} \equiv \hat{E}(Y_{it} \mid Z_{it}) = \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} Y_{js}K_{h,it,js}/\hat{f}_{it}$$

$$\hat{X}_{it} = \hat{E}(X_{it} \mid Z_{it}) = \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} X_{js}K_{h,it,js}/\hat{f}_{it}$$

$$\hat{W}_{it} = (\hat{W}_{1,it}, \hat{W}'_{2,it})'$$

且有

$$\hat{W}_{1,it} \equiv \hat{E}(Y_{i,t-1} \mid Z_{i,t-1}) = \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} Y_{j,s-1}K_{h,it-1,js}/\hat{f}_{i,t-1}$$

$$\hat{W}_{2,it} \equiv X_{2,it} - \hat{E}(X_{2,it} \mid Z_{i,t-1})$$
$$= X_{2,it} - \frac{1}{NT}\sum_{j=1}^{N}\sum_{s=1}^{T} X_{2,js}K_{h,it-1,js}/\hat{f}_{i,t-1}$$

其中 $K_{h,,it,js} = K_h((Z_{it} - Z_{js})/h)$ 是一个乘积核函数。

我们通过 $\hat{v}_{it} = X_{it} - \hat{X}_{it}$ 来估计 $v_{it} = X_{it} - E(X_{it} \mid Z_{it})$。那么,使用向量—矩阵符号,$\tilde{\alpha}_{IV}$ 的可行估计通过从式(19.63)中将

$$\xi_{it} \equiv E(Y_{it} \mid Z_{it}), \quad v_{it} = X_{it} - E(X_{it} \mid Z_{it})$$

以及 W_{it} 分别替换为核估计量 $\hat{\xi}_{it}$、$\hat{v}_{it} = X_{it} - \hat{X}_{it}$ 和 $\hat{W}_{it}$ 来获得,其矩阵形式为

$$\hat{\alpha}_{IV} = [\hat{v}'\hat{w}\hat{w}'\hat{v}]^{-1}\hat{v}'\hat{w}\hat{w}'(\mathcal{Y} - \hat{\xi}) \tag{19.71}$$

Baltagi 和 Li(2002)证明了 $\hat{\alpha}_{IV}$ 与式(19.65)中给出的 $\hat{\alpha}_{IV}$ 有相同的渐近分布。获得了 α 的一个 $\sqrt{N}$—致估计量后,可以基于下式来估计 $g(z)$:

$$Y_{it} - X_{it}'\hat{\alpha}_{IV} = g(Z_{it}) + u_{it} + X_{it}'(\alpha - \hat{\alpha}_{IV}) \equiv g(Z_{it}) + \varepsilon_{it}$$

其中 $\varepsilon_{it} = u_{it} + X_{it}'(\alpha - \hat{\alpha}_{IV})$。$g(z)$ 的一个非参数核估计量被给出为

599

$$\hat{g}(z) = \frac{1}{NT}\frac{\sum_i\sum_t (Y_{it} - X_{it}'\hat{\alpha})K_h\left(\frac{Z_{it} - z}{h}\right)}{\hat{f}(Z_{it})}$$

因为 $\hat{\alpha}_{IV} - \alpha = O_p(N^{-1/2})$,其快于通常的非参数收敛速度,很容易证明 $\hat{g}(z)$ 与 α 已知时有相同的渐近分布。因为这是一个相当标准的结果,在此,我们将不列出 $\hat{g}(z)$ 的渐近分布。

Lin 和 Ying(2001)提供了面板数据非参数和半参数估计的一般处理,Fan 和 Li(2004)考虑了半参数变系数面板数据模型,并且提出了同时选择显著性变量及估计未知参数的过程。

19.7　在半参数模型中检验序列相关以及个体效应

对于检验个体效应(或者零序列相关)的传统统计量通常都是典型地基于线性回归模型。这些检验过程对于回归函数形式误设是不稳健的。当基于一个线性模型拒绝没有个体效应(或者零序列相关)的原假定时,我们不清楚这个拒绝是由于真正的个体效应(或者误差序列相关)还是由于回归函数形式的误设。为了获得一个稳健的检验统计量,可以基于一个估计出来的半参数或者非参数回归模型来检验误差的序列相关。

Li 和 Hsiao(1998)考虑了在形如式(19.72)的一个半参数部分线性模型中

检验误差序列相关的问题。

$$Y_{it} = X_{it}'\alpha + g(Z_{it}) + u_{it} \tag{19.72}$$

其中 $g(\cdot)$ 的函数形式是未设定的，并且其中 X_{it} 可能包含 Y_{it} 的滞后值。Li 和 Hsiao 考虑的主要原假定是 $H_0:\rho = E[u_{it}u_{i,t-1}] = 0$（对应 $H_1:\rho \neq 0$）。他们基于 600 如上给定的部分线性模型的估计残差构建了一个检验统计量，并且他们推导了这一检验统计量的渐近分布。他们还提出了联合检验高阶序列相关和个体效应的方法。

Li 和 Hsiao(1998)考虑了一个密度加权统计量：

$$\rho_f = E[u_{it}f_{it}u_{i,t-1}f_{i,t-1}]$$

其中 $f_{it} = f(Z_{it})$ 是 Z_{it} 的概率密度函数。可以通过 $\hat{u}_{it}\hat{f}_{it}$ 来估计 $u_{it}f_{it}$，其中 $\hat{u}_{it} = Y_{it} - X_{it}'\hat{\alpha} - \hat{g}(Z_{it})$，$\hat{\alpha}$ 是第 18 章中讨论的 α 的半参数估计量，且 $\hat{g}(Z_{it}) = \hat{E}(Y_{it} \mid Z_{it}) - \hat{E}(X_{it} \mid Z_{it})'\hat{\alpha}$。检验统计量是 I_f 的样本对应值，即

$$I_{f,n} = N^{-1}\sum_i\sum_t \hat{u}_{it}\hat{u}_{i,t-1}\hat{f}_{it}\hat{f}_{i,t-1}$$

定理 19.2 在原假定下，误差 u_{it} 服从一个鞅差分过程，可以证明

$$\frac{\sqrt{N}I_{N,f}}{\sqrt{\hat{V}_I}} \xrightarrow{d} N(0,1)$$

其中

$$\hat{V}_I = N^{-1}\sum_i\sum_t[(\hat{u}_{i,t-1}\hat{f}_{i,t-1} - \hat{v}_{it}'\hat{f}_{it}\hat{B}^{-1}\hat{\Phi})\hat{u}_{it}\hat{f}_{it}]^2$$

$$\hat{B} = N^{-1}\sum_i\sum_t(X_{it} - \hat{X}_{it})(X_{it} - \hat{X}_{it})'\hat{f}_{it}^2, \hat{X}_{it} \equiv \hat{E}(X_{it} \mid Z_{it})$$

$$\hat{\Phi} = N^{-1}\sum_i\sum_t(Z_{it} - \hat{Z}_{it})\hat{f}_{it}\hat{u}_{i,t-1}\hat{f}_{i,t-1}$$

Li 和 Hsiao(1998)基于对所有 $t \neq s$，$E(u_{it}u_{is})$ 的一个密度加权方式，对于检验没有个体特定效应的原假定也提出了一个统计量。这个检验统计量被给定为

$$J_{N,f} = \frac{2}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s>t}^{T}\hat{u}_{it}\hat{u}_{is}\hat{f}_{it}\hat{f}_{is}$$

定理 19.3 在原假定下，误差 u_{it} 服从一个鞅差分进程，可以表明

$$\sqrt{N}J_{N,f}/\sqrt{\hat{V}_J} \xrightarrow{d} N(0,1)$$

601 其中

$$\hat{V}_J = \frac{2}{N(T-1)}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s>t}^{T}\sum_{s'>t}^{T}\hat{u}_{it}^2\hat{f}_{it}^2\{[\hat{u}_{is}\hat{f}_{is} - \hat{v}_{is}'\hat{f}_{is}\hat{B}^{-1}\hat{\Phi}_{t-s}]$$

$$\times [\hat{u}_{is}\hat{f}_{is'} - \hat{v}'_{is}\hat{f}_{is'}\hat{B}^{-1}\hat{\Phi}_{t-s'}]\}$$

$$\hat{\Phi}_{t-s} = N^{-1}\sum_{i}\sum_{s=t}^{T}\hat{v}_{it}\hat{f}_{it}\hat{u}_{i,t-s}\hat{f}_{i,t-s}$$

考虑一个形如式(19.73)的半参数动态面板数据模型

$$Y_{it} = X'_{it}\alpha + g(Z_{it}) + u_{it},\quad i = 1,2,\cdots,N,\ t = 1,2,\cdots,T \quad (19.73)$$

其中 X_{it} 是 $p\times 1$ 维的，α 是未知参数的一个 $p\times 1$ 维的向量，Z_{it} 是 q 维的，$g(\cdot)$ 是未设定的平滑函数。我们允许 X_{it} 包含 Y_{it} 的滞后值(比如，X_{it} 的第一个元素是 $Y_{i,t-1}$)，从而式(19.73)是一个半参数动态面板数据模型。我们假定在对 $s\leqslant t$，$E(u_{it}|Z_{is})=0$ 的意义上，Z_{it} 是弱外生的。同样，我们假定数据关于下标 i 是独立的，所以我们有 $\mathrm{var}(u) = I_N\otimes\Omega$，其中 $\Omega = \mathrm{var}(u_i)$，$u$ 是 $NT\times 1$ 维的，u_i 是 $T\times 1$ 维。我们考虑 T 固定、N 较大的情况，因此，所有渐近都是对 $N\to\infty$ 而言的。

我们允许误差 u_{it} 是序列相关的可能。例如，误差 u_{it} 可以有形如 $u_{it}=\mu_i+\nu_{it}$ 的一个单向误差成分设定，其中 μ_i 是一个随机个体效应，使得误差序列相关。

Kniesner 和 Li(2002)考虑了形如式(19.74)的一个半参数动态面板模型的估计

$$Y_{it} = Y_{i,t-1}\alpha + g(Z_{it}) + u_{it} \quad (19.74)$$

其中 Y_{it} 是个体 i 在时期 t 工作小时的对数，Z_{it} 包括 ln(工资)、教育、年龄、孩子数目等，Kniesner 和 Li 首先通过设定 $g(Z_{it}) = Z'_{it}\gamma$ 估计一个线性模型。他们基于普通最小二乘残差来检验没有个体效应的原假定，并且拒绝了原假定。若把这个结果作为误差真正序列相关或者个体效应的证据，我们需要使用某些工具估计方法来估计 α。然而，如上所述，检验结果实际上可能反映回归函数形式的误设。Kniesner 和 Li 接着估计了半参数模型(19.74)，并且应用了 Li 和 Hsiao(1998)的 $J_{N,f}$ 检验无个体效应，他们使用如上介绍的来自部分线性模型的估计残差，并且发现个体效应的存在没有相应的证据。他们进一步使用 12.1.1 节介绍的检验统计量(12.7)来检验是否 $g(Z) = Z'\gamma$，检验强烈地拒绝 $g(\cdot)$ 的线性函数形式。因此，他们的应用论证了函数形式的灵活使用如何可以帮助识别是否存在真正的个体效应或者是序列相关的误差，或者是否存在函数形式误设。 602

19.8 面板数据模型的序列估计

如我们在第 15 章讨论的，序列方法的一个优势就是其能容易地施加限制，如可以施加可加性。在本章余下的部分，我们关注面板数据模型的序列估计。

这些讨论主要来自 Ai 和 Li(2006)的工作。

19.8.1 可加性效应

我们考虑一个一般的可加性部分线性固定效应模型:

$$Y_{it} = X_{0it}'\alpha + \sum_{j=1}^{J} g_j(X_{jit}) + \mu_i + \nu_{it} \tag{19.75}$$

其中 $g_j(\cdot)$ 的函数形式是未设定的。

令 $q_j(\cdot) = (q_{j1}(\cdot), q_{j2}(\cdot), \cdots, q_{jk_j}(\cdot))'$ 表示已知序列基础函数,在下式的意义上能够很好地逼近 $g_j(\cdot)$,对某些 $\tau_j > 0$,有

$$g_j(\cdot) = q_j(\cdot)'\pi_j + O(k^{-\tau_j})$$

忽略逼近误差,参数向量 α 和系数向量 π_j 可以通过如下的非线性最小二乘估计量估计出来:

$$(\hat{\alpha}, \hat{\pi}) = \arg\min_{\alpha,\pi} \sum_{i=1}^{N}\sum_{t=2}^{T}\left[Y_{it} - Y_{i,t-1} - (X_{0it} - X_{0it-1})'\alpha - \sum_{j=1}^{J}[q_j(X_{jit}) - q(X_{jit-1})]'\pi_j\right]^2 \tag{19.76}$$

603 方程(19.76)是一个可加性部分线性模型的序列估计。$\hat{g}_j(\cdot) = q_j(\cdot)'\hat{\pi}_j$ 是 $g_j(\cdot)$ 的一个一致估计量,有通常非参数的收敛速度。$\hat{\alpha}$ 是 $\sqrt{N}$—一致且渐近正态分布的。$\hat{\alpha}$ 渐近方差的一个一致估计量可以通过把式(19.76)看成是一个参数回归模型而获得。更具体地,令 $\theta = (\alpha', \pi_1', \cdots, \pi_J')' \equiv (\alpha', \pi')'$,并且定义

$$l_i(\theta) = \sum_{t=1}^{T}\left[Y_{it} - Y_{it-1} - (X_{0it} - X_{0it-1})'\alpha \times - \sum_{j=1}^{J}[q(X_{jit}) - q(X_{jit-1})]'\pi_j\right]^2$$

$$\hat{V} = \left(\sum_{i=1}^{N}\frac{\partial^2 l_i(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\left(\frac{\partial l_i(\hat{\theta})}{\partial\theta}\frac{\partial l_i(\hat{\theta})}{\partial\theta'}\right)\left(\frac{\partial^2 l_i(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1} = \begin{bmatrix}\hat{V}_\alpha, & \hat{V}_{\alpha\pi}\\ \hat{V}_{\pi\alpha}, & \hat{V}_\pi\end{bmatrix}$$

注意对于 k_j(对所有 $j=1,\cdots,J$)的固定值,若 $q_j(\cdot)'\pi_j$ 是关于 $g_j(\cdot)$ 的正确设定,$\hat{V}_\alpha$ 将会是 $\hat{\theta}$ 方差的正确估计。但是,当 $k_j \to \infty$ 时,$\hat{V}$ 不是 $\hat{\delta}$ 的适当的估计方差,因为 $q_j(\cdot)'\pi_j$ 不是 $g_j(\cdot)$ 的一个 $\sqrt{N}$—一致估计。然而,$\hat{V}_\alpha$ 的左上角部

分总是给出 var($\hat{\alpha}$)的正确(即一致)估计量。

Ai 和 Chen(2003)的结论也可以被用来估计更一般的模型：

$$Y_{it} = v_0(X_{it},\alpha) + \sum_{j=1}^{J} g_j(v_j(X_{it},\alpha)) + \mu_i + \nu_{it} \tag{19.77}$$

其中 $v_j(\cdot)$，$j=1,\cdots,J$ 是已知形式的函数，$g_j(\cdot)$是未知的函数。令 $q_j(\cdot) = (q_{j1}(\cdot),q_{j2}(\cdot),\cdots,q_{jk_j}(\cdot))'$表示在下式意义上很好地逼近 $g_j(\cdot)$的一个已知序列，对某些 $\tau_j>0$，有

$$g_j(\cdot) = q_j(\cdot)'\pi_j + O(k_j^{-\tau_j})$$

忽略逼近误差，参数向量 α 和系数向量 π 可以通过非线性最小二乘估计量被估计出来：

604

$$(\hat{\alpha},\hat{\pi}) = \arg\min_{\alpha,\pi}\sum_{i=1}^{N}\sum_{t=2}^{T}\Big[Y_{it} - Y_{i,t-1} - v_0(X_{it},\alpha) + v_0(X_{it-1},\alpha)$$
$$- \sum_{j=1}^{J}[q_j(v_j(X_{it},\alpha)) - q_j(v_j(X_{i,t-1},\alpha))]'\pi_j\Big] \tag{19.78}$$

在类似于 Ai 和 Chen(2003)的条件下，$\hat{g}_j(\cdot) = q_j(\cdot)'\hat{\pi}_j$ 可以被证明是 $g_j(\cdot)$的一个一致估计量，有通常非参数的收敛速度($j=1,\cdots,J$)。$\hat{\alpha}$ 是$\sqrt{N}$—一致且渐近正态分布的。可以通过把式(19.78)看成是一个参数回归模型而获得 $\hat{\alpha}$ 渐近方差的一个一致估计量。更具体地，令 $\theta=(\alpha,\pi)$，并且定义

$$l_i(\theta) = \sum_{t=1}^{T}\Big[Y_{it} - Y_{i,t-1} - [v_0(X_{it},\alpha) - v_0(X_{it},\alpha)]$$
$$- \sum_{j=1}^{J}[q_j(v_j(X_{it},\alpha) - q_j(v_j(X_{it-1},\alpha))]'\pi_j]^2$$

$$\hat{V} = \left(\sum_{i=1}^{N}\frac{\partial^2 l_i(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}\left(\frac{\partial l_i(\hat{\theta})}{\partial\theta}\frac{\partial l_i(\hat{\theta})}{\partial\theta'}\right)\left(\frac{\partial^2 l_i(\hat{\theta})}{\partial\theta\partial\theta'}\right)^{-1}$$

$$= \begin{bmatrix}\hat{V}_\alpha, & \hat{V}_{\alpha\pi}\\ \hat{V}_{\pi\alpha}, & \hat{V}_\pi\end{bmatrix}$$

那么 $\hat{V}_\alpha$ 是 $\hat{\alpha}$ 的正确估计方差。

19.8.2　固定效应的其他表述

我们考虑一个具有固定效应的简单线性面板数据模型：

$$Y_{it} = X_{it}'\alpha + \mu_i + \nu_{it},\quad i=1,2,\cdots,N,\ t=1,2,\cdots,T \tag{19.79}$$

在本节我们没有进行一阶差分来消除固定效应 μ_i。相反地,我们假定存在

605 一些不随时间变化的变量 W_i,如 $E(\mu_i | X_i, W_i) = g(X_i, W_i)$,其中 $g(\cdot)$ 的函数形式是没有设定的。正如我们将简单地证明,这将导致一个半参数部分线性模型。我们将这种所谓的固定效应模型解释为一个随机效应 μ_i,使得 μ_i 和 X_i 的联合分布是未限制的。[①] 此外,假定回归元满足外生性条件 $E(\nu_i | X_i) = 0$,其中 $\nu_i = (\nu_{i1}, \cdots, \nu_{iT})$。在这个条件下,可知不随时间变化的回归元系数是不可识别的。因此,为了简便,我们假定 X_{it} 不包括时间不变回归元。我们用 W_i 来表示时间不变回归元,并且把外生性条件写为 $E(\nu_i | X_i, W_i) = 0$。时间不变回归元的系数 α 可以通过固定效应估计量 $\hat{\alpha}$ 被一致估计。固定效应估计量的优势是它不要求 (X_i, W_i) 和 μ_i 分布的参数化。它的劣势是它不能够估计时间不变回归元的效应。为了估计时间不变回归元的效应,应用研究者经常诉诸限制条件 $E(\mu_i | X_i, W_i) = W_i'\delta$ 并且应用随机效应估计量。当且仅当限制条件 $E(\mu_i | X_i, W_i) = W_i'\delta$ 满足时,随机效应估计量是一致的。

若 $E(\mu_i | X_i, W_i) \equiv g(X_i, W_i) \neq W_i'\delta$,那么可以简单地把 $g(\cdot, \cdot)$ 看成是模型中的一个未知非参数成分。通过这种方式,可以识别和估计时间不变回归元的效应。为了看出这一点,注意因为个体效应和回归元的联合分布是未限制的,函数 $g(X_i, W_i)$ 是未限制的。令 $\eta_i = \mu_i - g(X_i, W_i)$,那么 $E(\eta_i | X_i, W_i) = 0$,我们把模型(19.79)重写为

$$Y_i = X_i'\alpha + g(X_i, W_i) + u_{it}, \quad u_{it} = \eta_i + \nu_{it}, \quad E(u_{it} | X_i, W_i) = 0 \tag{19.80}$$

很明显,当 α 可识别时,未知的函数 $g(X_i, W_i) = E(Y_{it} - X_{it}'\alpha | X_i, W_i)$ 是可识别的,并且我们可以使用非参数序列方法来估计 α 和 $g(\cdot)$。为了说明这一点,令

$$p(x, w) = (p_1(x, w), p_2(x, w), \cdots, p_K(x, w))'$$

表示一个已知的基础序列(如样条或者幂级数函数),对一个 $K \times 1$ 向量 β 和一个标量 $\tau > 0$,满足

$$g(x, w) = p(x, w)'\beta + O(K^{-\tau})$$

606 忽略逼近误差 $O(K^{-\tau})$,我们通过最小二乘估计量来估计系数 α 和 β

$$(\hat{\alpha}, \hat{\beta}) = \arg\min_{\alpha, \beta} \sum_{i=1}^{N} \sum_{t=1}^{T} (Y_{it} - X_{it}'\alpha - p(X_i, W_i)'\beta)^2 \tag{19.81}$$

且未知函数 $g(\cdot)$ 通过 $\hat{g}(x, w) = p(x, w)'\hat{\beta}$ 来估计。$\hat{\alpha}$ 是 $\sqrt{N}$ 一致且渐近正态

① 也就是说,我们把固定效应看成随机效应,但是随机效应项 μ_i 可以和 $X_i = (X_{i1}, \cdots, X_{iT})$ 相关,其中相关的结构是未设定的。

分布的。$\hat{\alpha}$ 渐近方差的一个一致估计量可以通过将式(19.81)看成是一个参数(线性)回归模型,由其通常的方差估计量而轻易获得。

估计量 $\hat{g}(x,w)$ 可以被用来分析 w 的边际效应以及检验随机效应假定 $E(\mu_i|X_i,W_i)\neq W_i'\delta$。例如,$w$ 的平均边际效应是广义平滑参数 $\hat{g}(x,w)$ 的一个特殊情况,其在 Ai 和 Chen(2003)中被证明是渐近正态分布的。在第 12 章提出了对于假定 $g(x,w)=w'\delta$ 的一个一致检验。

19.9　非线性面板数据模型

线性面板数据模型的一个重要特征是观测的因变量关于个体效应是线性的,以上讨论的每个估计量都利用了这一点,因此个体效应可以通过简单的时间差分技术被消除。这个结构使得我们也能够在(X_i,W_i)均值独立于(η_i,ν_{it})的条件下估计条件均值 $E(\mu_i|X_i,W_i)$,其中 $\eta_i=\mu_i-E(\mu_i|X_i,W_i)$。然而,若观测的因变量关于个体效应是非线性的,个体效应不能够通过简单的时间差分技术被消除,并且 $E(\mu_i|X_i,W_i)$在均值独立条件下不能被一致估计。受限和定性因变量面板数据模型是两个重要的例子,其观测因变量关于个体效应是非线性的。在这些模型中,可能要求"非线性差分"技术来消除个体效应。本节的目的是综述这些在文献中已经被提及的非线性差分技术。我们回顾了面板数据审查回归模型(也被称为类型 1 Tobit 模型)、面板数据样本选择模型(也被称为类型 2 Tobit 模型,我们在第 10 章讨论过)以及面板数据离散选择模型。

我们将使用 $\mathbf{1}\{A\}$ 来表示一个示性函数,若事件 A 发生则取 1,否则取 0。　607
我们也将使用符号函数,$\mathrm{sgn}(A)$,若 A 是正的其取值为 1,若 A 是零其取值为 0,若 A 是负的其取值为 -1。对于每一个个体,我们有 $X_i=(X_{i1},\cdots,X_{iT})$, $Y_i=(Y_{i1},\cdots,Y_{iT})$, $u_i=(u_{i1},\cdots,u_{iT})$。最后,通过星号(*)标记的变量是不可观测的潜变量。

19.9.1　审查面板数据模型

我们从面板数据审查回归模型开始,其被给定为

$$\begin{aligned} Y_{it}^* &= X_{it}'\alpha+\mu_i+\nu_{it} \\ Y_{it} &= \max\{0,Y_{it}^*\} \end{aligned} \tag{19.82}$$

在这个模型中,潜在因变量 Y_{it}^* 关于个体效应 μ_i 是线性的,但观测的因变量 Y_{it}关于 μ_i 是非线性的,具有审查所引起的非线性。对观测因变量的简单时间差分不能够移除 μ_i。为了看出为什么简单时间差分不能够移除个体效应,对于任

何时期 t 以及真实值 α，记

$$Y_{it}-X_{it}'\alpha=\max\{Y_{it}^{*}-X_{it}'\alpha,-X_{it}'\alpha\}=\max\{\mu_i+\nu_{it},-X_{it}'\alpha\} \quad (19.83)$$

很明显，$Y_{it}-X_{it}'\alpha$ 是误差项 $\mu_i+\nu_{it}$ 的审查部分，其中 $-X_{it}'\alpha$ 是审查点。类似地，对于任何时期 s，

$$Y_{is}-X_{is}'\alpha=\max\{\mu_i+\nu_{is},-X_{is}'\alpha\} \quad (19.84)$$

是误差项 $\mu_i+\nu_{it}$ 的审查部分，其中 $-X_{is}'\alpha$ 作为审查点。应用简单的时间差分，我们得到

$$Y_{it}-X_{it}'\alpha-(Y_{is}-X_{is}'\alpha)=\max(\mu_i+\nu_{it},-X_{it}'\alpha)-\max(\mu_i+\nu_{is},-X_{is}'\alpha)$$

很明显，通过简单线性时间差分没有消除个体效应。

为了估计未知的参数 α，没必要在每一个数据点消除个体效应。若差分的误差项

$$\max(\mu_i+\nu_{it},-X_{it}'\alpha)-\max(\mu_i+\nu_{is},-X_{is}'\alpha)$$

608 有在给定回归元下的零条件均值，那么我们能够通过标准回归技术估计出 α。遗憾的是，当(i)条件于回归元和个体效应的误差项 ν_{it} 和 ν_{is} 不是同分布，或者(ii) 审查点 $-X_{it}'\alpha$ 和 $-X_{is}'\alpha$ 是不同的，则差分的误差项没有零条件均值。那么，为了获得 α 的一致估计，我们必须解决(i)和(ii)这两个问题。第一个问题通常通过施加如下的条件进行处理：

假定 19.1 误差项 ν_{it} 和 ν_{is} 条件于 (X_{it},X_{is},μ_i) 是同分布的。

第二个问题可以解决如下：通过人工审查观测因变量使得两个误差项在相同的审查点被审查，即 $\max\{-X_{it}'\alpha,-X_{is}'\alpha\}$。具体地，定义人工审查误差项为(下面第二个不等式利用了式(19.83))

$$\begin{aligned}e(Y_{it}-X_{it}'\alpha,X_{is}'\alpha)&\stackrel{\text{def}}{=}\max\{Y_{it}-X_{it}'\alpha,-X_{is}'\alpha\}\\&=\max\{\mu_i+\nu_{it},-X_{it}'\alpha,-X_{is}'\alpha\}\\e(Y_{is}-X_{is}'\alpha,X_{it}'\alpha)&=\max\{Y_{is}-X_{is}'\alpha,-X_{it}'\alpha\}\\&=\max\{\mu_i+\nu_{is},-X_{it}'\alpha,-X_{is}'\alpha\}\end{aligned}$$

接着从假定 19.1 可知 $e(Y_{it}-X_{it}'\alpha,X_{is}'\alpha)$ 和 $e(Y_{is}-X_{is}'\alpha,X_{it}'\alpha)$ 条件于回归元和个体效应是同分布的。这暗示着(μ_i 通过差分被消除了)

$$E[e(Y_{it}-X_{it}'\alpha,X_{is}'\alpha)-e(Y_{is}-X_{is}'\alpha,X_{it}'\alpha)\mid X_{it},X_{is}]=0 \quad (19.85)$$

因此，α 可以从条件矩约束(19.85)通过使用标准回归技术如 GMM 被一致估计。因为式(19.85)是通过观测因变量的一个非线性转换后应用简单的时间差分而获得的，我们把这种方法称为“非线性差分”。

尽管 α 的真实值可以从式(19.85)用 GMM 估计，但基于一个条件零均值

约束的估计比基于一个无条件零均值约束的估计更复杂。因此，一个更简单的方法是要找到凸目标函数，其一阶条件与式(19.85)隐含的无条件矩条件的某版本是相符的。例如，可能使用目标函数

$$A(\alpha) \stackrel{\text{def}}{=} E(r(Y_{it}, Y_{is}, (X_{it} - X_{is})'\alpha))$$

其中

$$r(Y_1, Y_2, \delta) = \begin{cases} Y_1^2 - \delta Y_1 - Y_1 Y_2, & \delta \leqslant -Y_2 \\ (Y_1 - Y_2 - \delta)^2, & -Y_2 < \delta < Y_1 \\ Y_2^2 + \delta Y_2 - Y_1 Y_2, & Y_1 \leqslant \delta \end{cases}$$

满足无条件矩约束

$$\frac{\partial A(\alpha)}{\partial \alpha} = E([e(Y_{it} - X_{it}'\alpha, X_{is}'\alpha) - e(Y_{is} - X_{is}'\alpha, X_{it}'\alpha)](X_{it} - X_{is})) = 0$$

很容易看出，对所有(Y_1, Y_2, δ)，$r(Y_1, Y_2, \delta) \geqslant 0$。因此，$\alpha$ 可以通过最小化$A(\alpha)$的样本对应物被估计出来：

$$\hat{\alpha} = \arg\min_{\alpha} \sum_{i=1}^{N} \sum_{t<s} r(Y_{it}, Y_{is}, (X_{it} - X_{is})'\alpha) \tag{19.86}$$

这就是 Honoré(1992)所建议的方法，他证明这个目标函数识别了 α，并且证明 $\hat{\alpha}$ 是$\sqrt{N}$一致且渐近正态分布的。可以通过通常的公式估计 $\hat{\alpha}$ 的渐近方差，给定为

$$\hat{V}_{\alpha} = \hat{A}^{-1}\hat{B}\hat{A}^{-1}$$

其中

$$\hat{A} = \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s<t}^{T} \frac{\partial^2 r(Y_{it}, Y_{is}, (X_{it} - X_{is})'\hat{\alpha})}{\partial \delta^2}(X_{it} - X_{is})(X_{it} - X_{is})'$$

$$\hat{B} = \sum_{i=1}^{N}\sum_{t}\sum_{s<t} D_{its} D_{its}'$$

$$D_{its} = \frac{\partial r(Y_{it}, Y_{is}, (X_{it} - X_{is})'\hat{\delta}}{\partial \alpha},(X_{it} - X_{is})$$

且$\dfrac{\partial^2 r(\cdot,\cdot,a)}{\partial \delta^2} = \dfrac{\partial^2 r(\cdot,\cdot,\delta)}{\partial \delta^2}\Big|_{\delta=a}$。

注意假定 19.1 也暗示着

$$E[\xi(e(Y_{it} - X_{it}'\alpha, X_{is}'\alpha)) - \xi(e(Y_{is} - X_{is}'\alpha, X_{it}'\alpha)) \mid X_{it}, X_{is}] = 0 \tag{19.87}$$

对于任何函数 $\xi(\cdot)$ 都成立。由此,被 ξ 标记的一类估计量可以从上面给出的条件矩约束通过 GMM 估计来获得;参见 Ai 和 Li(2006)关于如何基于式(19.87)来构建 α 的一致估计量。

610

在假定 19.1 下,上面描述的估计量可以被扩展到可加部分线性面板数据 Tobit 模型:

$$Y_{it} = \begin{cases} X'_{0,it}\alpha + \sum_{j=1}^{J} g_j(X_{j,it}) + \mu_i + \nu_{it}, \text{RHS} > 0 \\ 0, \qquad\qquad\qquad\qquad\qquad\qquad \text{其他} \end{cases}$$

其中 $g_j(\cdot)$ 是未知函数。出于识别的目的,我们假定未知函数满足对所有 j, $g_j(0)=0$。定义

$$e\left(Y_{it} - X'_{0,it}\alpha - \sum_{j=1}^{J} g_j(X_{j,it}), X'_{0,is}\alpha + \sum_{j=1}^{J} g_j(X_{j,is})\right)$$
$$= \max\left\{\mu_i + \nu_{it}, -X'_{0,it}\alpha - \sum_{j=1}^{J} g_j(X_{j,it}), -X'_{0,is}\alpha - \sum_{j=1}^{J} g_j(X_{j,is})\right\}$$

假定 19.1 暗示着

$$E\left(\left[e\left(Y_{it} - X'_{0,it}\alpha - \sum_{j=1}^{J} g_j(X_{j,it}), X'_{0,is}\alpha + \sum_{j=1}^{J} g_j(X_{j,is})\right)\right]\right.$$
$$\left. - \left[e\left(Y_{is} - X'_{0,is}\alpha - \sum_{j=1}^{J} g_j(X_{j,is}), X'_{0,it}\alpha + \sum_{j=1}^{J} g_j(X_{j,it})\right)\right] \middle| X_{it}, X_{is}\right)$$
$$= 0$$

按照 Ai 和 Chen(2003),我们通过线性筛 $p^{k_j}(\cdot)'\pi$ 来逼近 $p^k(\cdot)$,其中 $p^k(\cdot)$ 是逼近函数的一个向量。未知的参数向量 α 和筛系数 π 可以通过下式被估计:

$$(\hat{\alpha}_\xi, \hat{\pi}) = \arg\min_{\alpha,\pi} \sum_{i=1}^{N} \sum_{t<s} r\left[Y_{it}, Y_{is}, (X_{0,it} - X_{0,is})'\alpha\right.$$
$$\left. + \sum_{j=1}^{J} (p^{k_j}(X_{j,it}) - p^{k_j}(X_{j,is}))'\pi_j\right]$$

未知函数通过 $\hat{g}_j(\cdot) = p^{k_j}(\cdot)'\hat{\pi}_j$ 被估计出来。Ai 和 Li(2006)证明 $\hat{\alpha}$ 是 $\sqrt{N}$—一致且渐近正态分布的,并且 $\hat{g}_j(\cdot)$ 一致估计了 $g_j(\cdot)$。

611

假定 19.1 对于上面所有讨论的估计量是关键的条件。这个条件比 Tobit 模型正常所做的假定更弱。例如,它允许依赖数据以及误差项与解释变量的相依(即异方差)。但它仍然是受限的,它排除先决或者内生回归元。对于先决或者内生回归元的情况,条件矩约束(19.87)不成立。在这种情况下,上面描述的过程需要被修正。我们现在证明如何修正上面的过程来一致地估计模型的参

数。我们从先决回归元的情况开始。定义 $X_{it}=(X_{1it}',X_{2it}')'$ 并且相应地分解 $\alpha=(\alpha_1',\alpha_2')'$。假定 X_{1it} 是外生的，X_{2it} 是先决的（即包含滞后因变量）。假定 19.1 被修正如下：

假定 19.2　对任何 $t>s$，误差项 ν_{it} 和 ν_{is} 条件于 (X_{it},X_{is},μ_i) 是同分布的。

假定 19.2 弱于假定 19.1。为了看出为什么假定 19.2 允许先决回归元，假定目前 ν_{it} 独立于 $X_{1i}=(X_{1,i1},X_{1,i2}\cdots,X_{1,iT})$ 和 $X_{2i}^t=(X_{2,i1},X_{2,i2},\cdots,X_{2,it})$。那么，对任何 $s<t$，ν_{it} 和 ν_{is} 都独立于 X_{1i} 和 X_{2i}^s，并且只要 ν_{it} 和 ν_{is} 是同分布的，假定 19.2 被满足。

先决回归元的问题是审查点包含先决回归元，使得审查误差项 $e(Y_{it}-X_{it}'\alpha,X_{is}'\alpha)$ 和 $e(Y_{is}-X_{is}'\alpha,X_{it}'\alpha)$ 不是同分布的。这个问题的一个解决方式是从审查点去掉先决回归元。为了做到这一点，Honoré 和 Hu（1999）假定 $X_{2it}'\alpha_2\geqslant 0$ 以 1 的概率成立。接下来，定义

$$
\begin{aligned}
\varepsilon(Y_{it}-X_{it}'\alpha,X_{1,it}'\alpha_1,X_{is}'\alpha) &= \max\{Y_{it}-X_{it}'\alpha,\ -X_{1it}'\alpha_1,\ -X_{is}'\alpha\}\\
&= \max\{\mu_i+\nu_{it},\ -X_{1,it}'\alpha_1,\ -X_{is}'\alpha\}\\
\varepsilon(Y_{is}-X_{is}'\alpha,X_{1,it}'\alpha_1) &= \max\{Y_{is}-X_{is}'\alpha,\ -X_{1,it}'\alpha_1\}\\
&= \max\{\mu_i+u_{is},\ -X_{1,it}'\alpha_1,\ -X_{is}'\alpha\}
\end{aligned}
$$

假定 19.2 以及 $X_{2,it}'\alpha_2\geqslant 0$ 概率为 1 的假定暗示着给定 $(X_{1,it},X_{is})$ 下，误差项 $\varepsilon(Y_{it}-X_{it}'\alpha,X_{1,it}'\alpha_1,X_{is}'\alpha)$ 和 $\varepsilon(Y_{is}-X_{is}'\alpha,X_{1,it}'\alpha_1)$ 是同分布的。这也暗示着 612

$$
E(\varepsilon(Y_{it}-X_{it}'\alpha,X_{1,it}'\alpha_1,X_{is}'\alpha)-\varepsilon(Y_{is}-X_{is}'\alpha,X_{1,it}'\alpha_1)\mid X_{1,it},X_{is})=0 \tag{19.88}
$$

参数 α 现在可以从条件矩条件（19.88）通过 GMM 被估计出来。

对于内生回归元 $(X_{2,it},X_{2,is})$ 的情况，假定 19.2 不成立。令 (Z_{it},Z_{is}) 表示 $(X_{2,it},X_{2,is})$ 的工具变量。取决于对内生回归元所施加的约束，我们可以做出如下两个假定之一：

假定 19.3　条件于 $(X_{1,it},X_{1,is},Z_{it},Z_{is},\mu_i)$，$(\nu_{it},X_{2,it},X_{2,is})$ 和 $(\nu_{is},X_{2,it},X_{2,is})$ 两项是同分布的。

假定 19.4　误差项 ν_{it} 和 ν_{is} 条件于 $(X_{1,it},X_{1,is},Z_{it},Z_{is},\mu_i)$ 是同分布的。

在假定 19.3 下，$e(Y_{it}-X_{it}'\alpha,X_{is}'\alpha)$ 和 $e(Y_{is}-X_{is}'\alpha,X_{it}'\alpha)$ 条件于 $(X_{1,it},X_{1,is},Z_{it},Z_{is},\mu_i)$ 是同分布的。这也暗示着

$$
E(e(Y_{it}-X_{it}'\alpha,X_{is}'\alpha)-e(Y_{is}-X_{is}'\alpha,X_{it}'\alpha)\mid X_{1,it},X_{1,is},Z_{it},Z_{is})=0 \tag{19.89}
$$

因此,α 可以从条件矩约束(19.89)通过 GMM 被一致地估计出来。注意条件(19.89)不要求 $X'_{2,it}\alpha_2 \geqslant 0$ 以及 $X'_{2,is}\alpha_2 \geqslant 0$。在 $X'_{2,it}\alpha_2 \geqslant 0$ 概率为 1 以及 $X'_{2,is}\alpha_2 \geqslant 0$ 的条件下,我们可以从审查点去掉 $X_{2,is}$,且修正审查误差项如下:

$$v(Y_{it} - X'_{it}\alpha, X'_{1,it}\alpha_1, X'_{1,is}\alpha_1) = \max\{Y_{it} - X'_{it}\alpha, -X'_{1,it}\alpha_1, -X'_{1,is}\alpha_1\}$$
$$= \max\{\mu_i + \nu_{it}, -X'_{1,it}\alpha_1, -X'_{1,is}\alpha_1\}$$
$$v(Y_{is} - X'_{is}\alpha, X'_{1,it}\alpha_1, X'_{1,is}\alpha_1) = \max\{Y_{is} - X'_{is}\alpha, -X'_{1,it}\alpha_1, -X'_{1,is}\alpha_1\}$$
$$= \max\{\mu_i + \nu_{is}, -X'_{1,it}\alpha_1, -X'_{1,is}\alpha_1\}$$

假定 19.4 意味着

$$E\{[v(Y_{it} - X'_{it}\alpha, X'_{1,it}\alpha_1, X'_{1,is}\alpha_1) - v(Y_{is} - X'_{is}\alpha, X'_{1,it}\alpha_1, X'_{1,is}\alpha_1)] \mid X_{1,it}, X_{1,is}, Z_{it}, Z_{is}\} = 0 \tag{19.90}$$

613 再次,α 可以从条件矩约束(19.90)通过 GMM 被一致地估计出来。

上面描述的观点可以很容易地被应用到如下动态潜在因变量面板数据模型:

$$\begin{aligned} Y^*_{it} &= \rho Y^*_{it-1} + X'_{it}\alpha + \mu_i + \nu_{it} \\ Y_{it} &= \max\{0, Y^*_{it}\} \end{aligned} \tag{19.91}$$

为了说明这一点,假定回归元 X_{it} 是严格外生的。注意这个模型不同于其他的 Tobit 模型,因为滞后的潜在因变量是不可观测的。首先,我们选择一个子样本,对某些 $s<t$, $Y_{is-1}>0$。那么,

$$Y_{is} - \rho Y_{is-1} - X'_{is}\alpha = \max\{\mu_i + \nu_{is}, -\rho Y_{is-1} - X'_{is}\alpha\}$$
$$Y_{it} - \rho Y^*_{it-1} - X'_{it}\alpha = \max\{\mu_i + \nu_{it}, -\rho Y^*_{it-1} - X'_{it}\alpha\}$$

假定 Y_{is-1} 独立于 ν_{it} 和 ν_{is} $(s<t)$ 是合理的。另一方面,Y^*_{it-1},从而 Y_{it-1},不独立于 ν_{is}。假定 $\rho>0$,那么条件于 $Y_{it-1}>0$,记

$$e(Y_{it} - \rho Y^*_{it-1} - X'_{it}\alpha, \rho Y_{is-1} + X'_{is}\alpha, X'_{it}\alpha)$$
$$= \max\{Y_{it} - \rho Y^*_{it-1} - X'_{it}\alpha, -\rho Y_{is-1} - X'_{is}\alpha, -X'_{it}\alpha\}$$
$$= \max\{\mu_i + \nu_{it}, -\rho Y_{is-1} - X'_{is}\alpha, -X'_{it}\alpha\}$$
$$e(Y_{is} - \rho Y_{is-1} - X'_{is}\alpha, X'_{it}\alpha) = \max\{Y_{is} - \rho Y_{is-1} - X'_{is}\alpha, -X'_{it}\alpha\}$$
$$= \max\{\mu_i + \nu_{is}, -\rho Y_{is-1} - X'_{is}\alpha, -X'_{it}\alpha\}$$

注意,条件于 $Y_{it-1}>0$,以下两项

$$\max\{\mu_i + \nu_{it}, -\rho Y_{is-1} - X'_{is}\alpha, -X'_{it}\alpha\}$$
$$\max\{\mu_i + \nu_{is}, -\rho Y_{is-1} - X'_{is}\alpha, -X'_{it}\alpha\}$$

不是同分布的。所以，为了恢复对称性，我们要求 ν_{it} 和 ν_{is} 满足相同的约束。例如，我们可以要求它们满足

$$\min\{\mu_i + \nu_{it-1}, \mu_i + \nu_{it}\} \geqslant \max\{-\rho Y_{it-2} - X'_{i2-1}\alpha, \ -X'_{it}\alpha\} \quad (19.92)$$

因此

$$Y^*_{it-1} \geqslant \max\{0, \rho Y_{it-2} + X'_{it-1}\alpha - X'_{it}\alpha\} > 0$$

$$Y^*_{it} \geqslant \max\{\rho Y_{it-1} + X'_{it}\alpha - \rho Y_{it-2} - X'_{it-1}\alpha, \rho Y_{it-1}\} > 0$$

假定 19.5　对任何 t，误差项 ν_{it} 和 ν_{it-1} 条件于 $(Y_{it-2} > 0, X_{it}, X_{is}, \mu_i)$ 是同分布的。

在假定 19.5 下，并条件于

614

$$\mathcal{A}_{its} = \{Y_{it-2} > 0, Y_{it-1} \geqslant \max\{0, \rho Y_{it-2} + X'_{it-1}\alpha - X'_{it}\alpha\}\}$$

以及

$$Y_{it} \equiv \max\{\rho Y_{it-1} + X'_{it}\alpha - \rho Y_{it-2} - X'_{it-1}\alpha, \rho Y_{it-1}\}$$

我们有以下两项

$$e(Y_{it} - \rho Y_{it-1} - X'_{it}\alpha, \rho Y_{it-2} + X'_{it-1}\alpha, X'_{it}\alpha)$$

$$e(Y_{it-1} - \rho Y_{it-2} - X'_{it-1}\alpha, \rho Y_{it-2} + X'_{it-1}\alpha, X'_{it}\alpha)$$

是同分布的。

这导致了下面的条件矩约束：

$$E\{\mathbf{1}(\mathcal{A}_{its})[e(Y_{it-1} - \rho Y_{it-2} - X'_{it-1}\alpha, X'_{it}\alpha) - e(Y_{it} - \rho Y_{it-1} - X'_{it}\alpha, \rho Y_{it-2} + X'_{it-1}\alpha, X'_{it}\alpha)] \mid X_{it}, X_{is}\} = 0$$

参数 α 现在可以从上面的条件矩约束被估计出来，估计量的渐近性质可以像 Hu(2002)那样被完全推导出来。

19.9.2　离散选择面板数据模型

Tobit 模型提出的非线性差分方法的一个关键方面是潜在因变量是部分可观测的，使得可以使用修剪来恢复误差分布的对称性。而这种技术对于面板数据离散选择模型是不起作用的，因为潜在因变量不可观测，从而必须开发一种新的非线性差分方法。在本节，我们回顾这些新的非线性差分技术。我们从面板数据二元选择模型开始，给定为

$$Y_{it} = \mathbf{1}\{X'_{it}\alpha + \mu_i + \nu_{it} > 0\}, \quad i = 1,2,\cdots,N, \ t = 1,2,\cdots,T \quad (19.93)$$

其中 X_{it} 是随时间变化的解释变量向量，μ_i 是个体特定截距，ν_{it} 是误差项。注意对于任何两时期 t 和 s，简单的时间差分得到

$$Y_{it} - Y_{is} = \mathbf{1}\{X'_{it}\alpha + \mu_i + \nu_{it} > 0\} - \mathbf{1}\{X'_{is}\alpha + \mu_i + \nu_{is} > 0\}$$

615 取期望,我们得到

$$E(Y_{it}-Y_{is}\mid X_{it},X_{is},\mu_i)=\mathrm{P}(\nu_{it}>-X_{it}'\alpha-\mu_i\mid X_{it},X_{is},\mu_i)$$
$$-\mathrm{P}(\nu_{is}>-X_{is}'\alpha-\mu_i\mid X_{it},X_{is},\mu_i)$$

很明显,简单的时间差分没有消除个体效应,除非 $X_{it}'\alpha=X_{is}'\alpha$。Manski (1987)观察到若右边差分的概率有与 $X_{it}'\alpha-X_{is}'\alpha$ 相同的符号,则 $Y_{it}-Y_{is}$ 与 $\mathrm{sgn}(X_{it}'\alpha-X_{is}'\alpha)$ 是正相关的。基于这个观察,Manski 提出一个最大得分估计量,其最大化 s 样本相关性,被给出为

$$\hat{\alpha}=\arg\max_{\alpha}\sum_{i=1}^{N}\sum_{s<t}(Y_{it}-Y_{is})\times\mathrm{sgn}((X_{it}-X_{is})'\alpha)$$

很明显,Manski 的估计量通过符号函数被定义在差分数据上。因为符号函数是非线性的,我们将这种方法也称为非线性差分。

为了确保差分后的概率有与 $X_{it}'\alpha-X_{is}'\alpha$ 相同的符号,我们施加如下的条件:

假定 19.6 对任何 $s<t$,误差项 ν_{it} 和 ν_{is} 条件于 (X_{it},X_{is},μ_i) 是同分布的。

如同针对 Tobit 模型的假定 19.1,假定 19.6 是较弱的。它没有要求设定误差分布,并且它允许依赖数据和异方差。在一些附加条件和对模型参数的比例正规化(scale normalization)下,Manski(1987)证明了最大得分估计量的一致性。然而,他的估计量既不是 $\sqrt{N}$ 一致也不是渐近正态分布的。他的估计量的非渐近正态分布的产生是由于非平滑的目标函数。若应用 Horowitz(1992)建议的平滑技术,得到的估计量可以被证明是渐近正态分布的,虽然收敛速度仍低于 $\sqrt{N}$;具体参见 Kyriazidou(1997)以及 Charlier, Melenberg 和 van Soest(1995)。

可以直接将 Manski(1987)的思路扩展到如下的非参数面板数据模型:

$$Y_{it}=\mathbf{1}\{g(X_{it})+\mu_i+\nu_{it}>0\},\quad i=1,2,\cdots,N,\ t=1,2,\cdots,T \tag{19.94}$$

616 其中 $g(\cdot)$ 的形式未知。令下式表示逼近于 $g(x)$ 的一个非参数序列函数

$$p^k(x)'\pi=p_1(x)\pi_1+\cdots+p_k(x)\pi_k$$

令

$$\hat{\pi}=\arg\max_{\pi}\sum_{i=1}^{N}\sum_{s<t}(Y_{it}-Y_{is})\times\mathrm{sgn}((p^k(X_{it})-p^k(X_{is}))'\pi)$$

且令 $\hat{g}(x)=p^k(x)'\hat{\pi}$。然后,使用 Shen(1997)和 Manski(1987)发展的技术,$\hat{g}(x)$ 可以被证明在 sup 和 L_2 范数下是一致的。

回到式(19.93),如同假定 19.1 的情况,假定 19.5 排除先决解释变量如滞后因变量的情况。若允许有先决解释变量,则 Honoré 和 Kyriazidou(2000)使用

的技巧可以在这里被用来估计模型(19.93)。具体地,分解 $X_{it}'\alpha = X_{1it}'\alpha_1 + X_{2it}'\alpha_2$,并且假定 X_{2it}是先决解释变量,考虑三个时期 $r < s < t$。Honoré 和 Kyriazidou 提供的见解得到如下的估计量:

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{i=1}^{N} \sum_{r<s<t} K\left(\frac{X_{1is} - X_{1it}}{h_N}\right)(Y_{is} - Y_{ir})$$

$$\times \operatorname{sgn}((X_{1is} - X_{1ir})'\alpha_1 + (X_{2it} - X_{2ir})'\alpha_2)$$

其中 $K(\cdot)$表示乘积核函数。Honoré 和 Kyriazidou 证明了这个估计量的一致性。

对于具有 $h(x) = g_1(X_1) + g_2(X_2)$的非参数模型(19.94),令下式表示对 $g_1(X_1)$的逼近:

$$p_1^{k_1}(X_1)'\pi_1 = p_{11}(X_1)\pi_{11} + \cdots + p_{1k_1}(X_1)\pi_{1k_1}$$

令下式表示对 $g_2(X_2)$的逼近:

$$p_2^{k_2}(X_2)'\pi_2 = p_{21}(X_2)\pi_{21} + \cdots + p_{2k_2}(X_2)\pi_{2k_2}$$

Honoré 和 Kyriazidou(2000)的思想又可以得到如下的估计量:

$$\hat{\pi} = \arg\max_{\alpha} \sum_{i=1}^{N} \sum_{r<s<t} K_{is,it}(Y_{is} - Y_{ir})$$

$$\times \operatorname{sgn}[(p_1^{k_1}(X_{1is}) - p_1^{k_1}(X_{1ir}))'\pi_1$$

$$+ (p_2^{k_2}(X_{2it}) - p_2^{k_2}(X_{2ir}))'\pi_2]$$

其中 $K_{is,it} = K((X_{1is} - X_{1it})/h_N)$。可以证明 $\hat{g}_j(\cdot) = p_j^{k_j}(\cdot)'\hat{\pi}_j$ 一致地估计了 $g_j(\cdot)$。

Manski(1987)的非线性差分方法在一定意义上是一般性的,即与二元选择模型正常要求的条件相比,它要求的条件弱很多。这种一般性的代价是估计量既不是$\sqrt{N}$一致也不是渐近正态分布的。因此,一个很自然的问题是,我们是否可以通过施加一个更强的条件来达到$\sqrt{N}$一致和渐近正态分布。Andersen(1970)通过考虑模型(19.93)的一个 Logit 版本回答了这个问题。他的非线性差分思想基于条件最大似然方法。令

617

$$\gamma_i = \sum_{t=1}^{T} Y_{it}$$

其取值在$\{0,1,\cdots,T\}$,因为 $Y_{it} \in \{0,1\}$。同样令 γ_{it}取值在$\{0,1\}$。那么,Andersen 建议通过下式估计 α:

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{i=1}^{N} \ln \left[\frac{\exp\left(\sum_{t=1}^{T} Y_{it} X_{it}' \alpha\right)}{\sum_{\gamma_{i1}+\cdots+\gamma_{iT}=\gamma_i} \exp\left(\sum_{t=1}^{T} \gamma_{it} X_{it}' \alpha\right)} \right]$$

其中 $\sum_{\gamma_{i1}+\cdots+\gamma_{iT}=\gamma_i}$ 表示($\gamma_{i1},\cdots,\gamma_{iT}$)所有可能组合的总和,有 $\sum_{t=1}^{T} \gamma_{it} = \gamma_i$。

Andersen(1970)证明条件最大似然方法是$\sqrt{N}$一致和渐近正态分布的。这个条件可以很容易地扩展到式(19.94)的 Logit 版本,其中

$$\hat{\pi} = \arg\max_{\alpha} \sum_{i=1}^{N} \ln \left[\frac{\exp\left(\sum_{t=1}^{T} Y_{it} p^k (X_{it})' \pi\right)}{\sum_{\gamma_{i1}+\cdots+\gamma_{iT}=\gamma_i} \exp\left(\sum_{t=1}^{T} \gamma_{it} p^k (X_{it}) \pi\right)} \right]$$

且 $\hat{g}(x) = p^k(x)'\hat{\pi}$。Shen(1997)准确地证明了 $\hat{g}(x)$ 平滑函数的一致和渐近正态性。

为了在面板数据 Logit 模型中允许有先决解释变量, Honoré 和 Kyriazidou (2000)要求至少有三个时期并且提出了如下的估计量:

$$\hat{\alpha} = \arg\max_{\alpha} \sum_{i=1}^{N} \sum_{r<s<t} \mathbf{1}\{Y_{ir} + Y_{is} = 1\} K_{is,it}$$
$$\times \ln\left(\frac{[\exp((X_{1ir} - X_{1is})'\alpha_1 + (X_{2ir} - X_{2it})'\alpha_2)]^{Y_{ir}}}{1 + \exp((X_{1ir} - X_{1is})'\alpha_1 + (X_{2ir} - X_{2it})'\alpha_2)} \right)$$

这个估计量有通常非参数的收敛速率。

总之,我们知道最大得分估计量不是$\sqrt{N}$一致的,但它对误差项的分布只施加了很弱的约束。另一方面,条件最大似然估计量是$\sqrt{N}$一致的,但要求对误差项的分布施加一个非常强的假定。很自然地会问是否存在对误差分布的其他约束,使得对于模型的参数是$\sqrt{N}$一致的估计量。遗憾的是, Chamberlain(1993)对于这个问题给出了令人惊讶的消极答案。他证明即使误差是独立同分布且独立于解释变量和个体效应,模型的参数也仅在 Logit 的情形下可以被$\sqrt{N}$一致估计。[①] 因此,为了得到对于非 Logit 模型的一个$\sqrt{N}$一致估计量,必须对解释变量和个体效应之间的相关性施加额外假定。Honoré 和 Lewbel(2002)要求存在一个"特别的回归元",它是连续的、有有限的支撑且独立于个体效应和误差项。

618

① 当参数没有被识别时,如何在动态离散选择面板数据模型中对参数规定界限,参见 Honoré 和 Tamer(2006)。

这个特殊的回归元的作用是将个体效应从非线性函数中移除。具体地，记

$$X_{it}'\alpha = X_{1it} + X_{2it}'\alpha_1$$

其中 X_{1it} 是一个特殊的回归元，出于可识别目的其系数标准化为 1，X_{2it} 表示先决回归元。对于任何两期 $r>s$，令 Z_{is} 表示直到 s 时期的由所有先决变量组成的工具变量。在一系列正则性条件下，Honoré 和 Lewbel 证明

$$E\left(\frac{Z_{is}(Y_{it}-\mathbf{1}\{X_{1it}>0\})}{f_t(X_{1it}\mid X_{2it},Z_{is})}\right) = Z_{is}X_{2it}'\alpha_1 + E(Z_{is}\mu_i), \quad t=r,s$$

其中 f_t 表示允许随时间变化的条件密度。通过时期 r 和 s 之间简单的时间差分现在可以消除个体效应，参数可以通过简单的工具变量方法来估计。具体参见 Honoré 和 Lewbel。

19.10　证　明

19.10.1　定理 19.1 的证明

因为这个证明类似于第 2 章中定理 2.2 的证明，这里我们仅仅提供一个概要。

证明：我们记

$$\hat{g}(x) - \hat{g}(x) = \hat{m}(x)/\hat{f}(x)$$

其中

$$\hat{m}(x) = [\hat{g}(x) - g(x)]\hat{f}(x)$$

我们进一步记

$$\hat{m}(x) = \hat{m}_1(x) + \hat{m}_2(x)$$

其中

$$\hat{m}_1(x) = \sum_i\sum_t (g(X_{it}) - g(x))K_{h,it,x}/\mathrm{NT}$$

$$\hat{m}_2(x) = \sum_i\sum_t u_{it}K_{h,it,x}/\mathrm{NT}$$

我们首先考虑 $\hat{m}_2(x)$。很明显，$E(\hat{m}_2(x))=0$，因为 $E(u_{it}\mid X_{it})=0$。为了评价 $\mathrm{var}(\hat{m}_2(x))$，我们考虑 N 和 T 都很大的情况，并且我们假定 $u_{it}=\mu_i+\nu_{it}$，其中 μ_i 是独立同分布 $(0,\sigma_\mu^2)$，ν_{it} 是独立同分布 $(0,\sigma_\nu^2)$，并且 μ_i 和 ν_{it} 是互相独立的。那么我们有

$$
\begin{aligned}
\operatorname{var}(\hat{m}_2(x)) &= E[(\hat{m}_2(x))^2] \\
&= (\mathrm{NT})^{-2}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{T}E[u_{it}u_{is}K_{h,it,x}K_{h,is,x}] \\
&= (\mathrm{NT}^2)^{-1}\left\{\mathrm{TE}(u_{it}^2K_{h,it,x}^2)+\sum_t\sum_{s\neq t}E(\mu_i^2K_{h,it,x}K_{h,is,x})\right. \\
&\quad \left.+\sum_t\sum_{s\neq t}E(\nu_{it}\nu_{is}K_{h,it,x}K_{h,is,x})\right\} \\
&\equiv m_{2,1}(x)+m_{2,2}(x)+m_{2,3}(x)
\end{aligned}
$$

现在，

$$
\begin{aligned}
m_{2,1}(x) &= (NT)^{-1}E[u_{it}^2K_{h,it,x}^2] = (NT)^{-1}E[\sigma^2(X_{it})K_{h,it,x}^2] \\
&= (NTH_q)^{-1}[\kappa^q\sigma^2(x)f(x)+o(1)]
\end{aligned}
$$

使用第18章的引理18.1，我们有

$$
\begin{aligned}
M_N &= E[\,|K_{h,it,x}K_{h,is,x}|^{1+\delta}] \\
&= \int |K_{h,it,x}K_{h,is,x}|^{1+\delta}f_{ts}(x_{it},x_{is})\,\mathrm{d}x_{it}\mathrm{d}x_{is} \\
&= H_q^{-2\delta}\int |K(u)K(v)|^{1+\delta}f_{ts}(x+hu,x+hv)\,\mathrm{d}u\mathrm{d}v \\
&\leqslant C_1H_q^{-2\delta}
\end{aligned}
$$

620 同样，

$$
\begin{aligned}
|m_{2,2}(x)| &= (NT^2)^{-1}\left|\sum_t\sum_{s\neq t}E(\mu_i^2K_{h,it,x}K_{h,is,x})\right| \\
&\leqslant C_1(NT^2)^{-1}\sum_t\sum_{s\neq t}E[(K_{h,it,x}K_{h,is,x})] \\
&\leqslant C_1(NT^2)^{-1}\left\{\sum_t\sum_{s\neq t}E(K_{h,it,x})E(K_{h,is,x})\right. \\
&\quad \left.+\sum_t\sum_{s\neq t}M_N^{\delta/(1+\delta)}\beta_{t-s}^{\delta/(1+\delta)}\right\} \\
&\leqslant C_2N^{-1}f(x)^2+C_3N^{-1}H_q^{-2/(1+\delta)}\sum_{\tau=1}^{\infty}\beta_t^{\delta/(1+\delta)} \\
&= O(N^{-1})+O((NH_q^{2/(1+\delta)})^{-1}) = o(1)
\end{aligned}
$$

因为 $TH_q=o(1)$ 以及 $0<\delta<1$。

类似地，我们可以证明

$$
|m_{2,3}|\leqslant C_3N^{-1}H_q^{-2/(1+\delta)}\sum_{\tau=1}^{\infty}\beta_\tau^{\delta/(1+\delta)} = o(1)
$$

综合以上我们已经证明了

$$\mathrm{var}(\hat{m}_2(x)) = \frac{1}{NTH_q}\kappa^q\sigma^2(x)f(x)[1+o(1)] \tag{19.95}$$

通过第2章中完全相同的证明,我们可以很容易地证明

$$E[\hat{m}_1(x)] = \sum_{s=1}^{q} h_s^2 B_s(x) + o_p\left(\sum_{s=1}^{q} h_s^2\right) \tag{19.96}$$

通过使用β混合不等式,像我们推导式(19.95)那样,我们可以证明

$$\mathrm{var}(\hat{m}_1(x)) = O\left(\sum_{s=1}^{q} h_s^2 (NTH^q)^{-1}\right)$$

因此,$\hat{m}(x)$的方差首项和偏误首项在式(19.95)和式(19.96)中给出,其与第2章中考虑的独立数据情形有相同的形式。

类似地,可以证明

621

$$\mathrm{MSE}[\hat{f}(x)] = O\left(\sum_{s=1}^{q} h_s^2 + (NTH_q)^{-1}\right)$$

其暗示着$\hat{f}(x) - f(x) = o_p(1)$。因此(中心极限定理成立的条件也可以被证实)

$$\begin{aligned} &\sqrt{NTH_q}\left[\hat{g}(x) - g(x) - \sum_{s=1}^{q} h_s B_s(x)\right] \\ &\sim \sqrt{NTH_q}\left[\hat{m}(x) - m(x) - \sum_{s=1}^{q} h_s B_s(x)\right] \Big/ f(x) \\ &\to N(0, \kappa^q\sigma^2(x)/f(x) \text{ 依分布收敛} \end{aligned}$$

□

19.10.2 Wang 估计量均方误差首项的计算

在这一节我们给出计算 Wang(2003)估计量$\hat{g}(z)$偏误首项和方差首项的大体证明。我们首先证明一个引理。

引理 19.1 令$\hat{g}_{[l]}(z)$表示$g(z)$的第l步估计量,那么

$$\begin{aligned} \hat{g}_{[l]}(z) - g(z) &= \frac{\kappa_2}{2}\sum_{r=1}^{q} h_r^2 g_{rr}(z) + \frac{1}{N\Omega(z)}\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it}, z)\varepsilon_{it} \\ &+ \frac{1}{N\Omega(z)}\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{s\neq t}^{T} K_h(Z_{it}, z)\ \mathcal{L}_{i,tsg}(\cdot)\{\hat{g}_{[l-1]}(z) - g(z)\} + o_p(\eta_N) \end{aligned} \tag{19.97}$$

其中

$$\kappa_2 = \int k(v) v^2 \mathrm{d}v$$

$$\varepsilon_{it} = \mathcal{L}_{i,tg}\{Y_i, g(Z_{i1}), \cdots, g(Z_{iT})\} = c_t'\Sigma^{-1}u_i = \sum_{s=1}^{T}\sigma^{ts}u_{is}$$

且 $\eta_N = \sum_{r=1}^{q} h_r^2 + (Nh_1 \cdots h_q)^{-1/2}$。

622 证明：对 $(\hat{g}_{[l]}(z), \hat{g}_{[l]}^{(1)}(z))$ 在 $(g(z), g^{(1)}(z)$ 处) 使用一阶方程(19.10)的一个 Taylor 展开式，我们有

$$0 = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it}, z) G_{it}(z,h)\ \mathcal{L}_{i,tg}(\cdot)$$
$$+ N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it}, z) G_{it}(z,h) G_{it}(z,h)'\ \mathcal{L}_{i,ttg}(\cdot)\begin{bmatrix}\hat{g}_{[l]}(z) - g(z)\\ \hat{g}_{[l]}^{(1)}(z) - g^{(1)}(z)\end{bmatrix}$$
$$+ o_p(\eta_N) \tag{19.98}$$

其中变量 $(\cdot)$ 被给出为

$$\{Y_i, \hat{g}_{[l-1]}(Z_{i1}), \cdots, g(z) + [(Z_{it} - z)/h]'g^{(1)}(z), \cdots, \hat{g}_{[l-1]}(Z_{iT})\}$$

很容易证明

$$N^{-1}\sum_{i=1}^{N} K_h(Z_{it}, z)\ \mathcal{L}_{i,ttg} G_{it} G_{it}' \sim E[K_h(Z_{it}, z)\ \mathcal{L}_{i,ttg} G_{it} G_{it}']$$

收敛于 $-\Omega(z)\begin{bmatrix}1, & 0\\ 0, & \kappa_2 I_q\end{bmatrix}$，其中 $\kappa_2 = \int k(v)v^2\mathrm{d}v$。因此式(19.98)导致

$$\Omega(z)(\hat{g}_{[l]}(z) - g(z)) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it}, z)\ \mathcal{L}_{i,tg}(\cdot) + o_p(N^{-1/2})$$
$$\equiv A_N + o_p(\eta_N)$$

其中

$$A_N(z) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it}, z)\ \mathcal{L}_{i,tg}(\cdot) = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it}, z)$$
$$\times \left[\sigma^{tt}(Y_{it} - g(z) - [(Z_{it} - z)/h]'g^{(1)}(z)) + \sum_{s \neq t}^{T}\sigma^{ts}(Y_{is} - \hat{g}_{[l-1]}(Z_{is}))\right]$$

我们分解 A_N 为 $A_N = A_{1N} + A_{2N}$，其中 A_{1N} 是从 A_N 中将 $\hat{g}_{[l-1]}(\cdot)$ 替换为

623 $g(\cdot)$ 而获得的，$A_{2N} = A_N - A_{1N}$。注意 $Y_{is} - g(Z_{is}) = u_{is}$，我们有

$$A_{1N} = \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it}, z)[\sigma^{tt}(g(Z_{it}) - g(z) - [(Z_{it} - z)/h]'g^{(1)}(z))]$$
$$+ \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T} K_h(Z_{it}, z)\sum_{s=1}^{T}\sigma^{ts}u_{is}$$
$$\equiv A_{1N1} + A_{1N2} \tag{19.99}$$

其中 A_{1N1} 和 A_{1N2} 的定义应该是显而易见的。

在 A_{1N1} 中将 $g(Z_{it})$ 在 z 处 Taylor 展开,得到

$$A_{1N1} = \frac{1}{2N}\sum_{i=1}^{N}\sum_{t=1}^{T}K_h(Z_{it},z)\sigma^{tt}[(Z_{it}-z)]g^{(2)}(z)[(Z_{it}-z)]' + o_p\left(\sum_{s=1} h_r^2\right) \quad (19.100)$$

可以证明, $A_{1N1}=E(A_{1N1})+o_p(\eta_N)$,并且(见习题 19.6)

$$E(A_{1N1}) = \frac{\kappa_2}{2}\Omega(z)\sum_{r=1}^{q}h_r^2 g_{rr}(z) + o\left(\sum_{s=1}^{q}h_r^2\right) \quad (19.101)$$

其中 $g_{rr}(z)=\partial^2 g(z)/\partial z_r^2(r=1,\cdots,q)$。

接下来,

$$\begin{aligned} A_{2N} &= A_N - A_{1N} \\ &= \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}K_h(Z_{it},z)\sum_{s=1}^{T}\sigma^{ts}[\hat{g}_{[l-1]}(Z_{is})-g(Z_{it})] \end{aligned} \quad (19.102)$$

这就完成了引理 19.1 的证明。 □

$\hat{g}(z)-g(z)$ 的方差首项来自 $\Omega(z)^{-1}A_{1N2}$。它的零均值和渐近方差,被给定为(见习题 19.2)

$$\Omega(z)^{-2}\mathrm{var}(A_{1N2}) = \frac{\kappa^q}{\Omega(z)}(Nh_1\cdots h_q)^{-1} + o((Nh_1\cdots h_q)^{-1}) \quad (19.103)$$

重复应用引理 19.1,导致 $\hat{g}(z)-g(z)$ 的偏误首项被给定为式(19.14)。若仅迭代一步,那么偏误被给定为 A_{1N1} 和 A_{2N}, $l=1$,而 $\hat{g}_{[l-1]}(Z_{is})=\hat{g}_{[0]}(Z_{is})$ 就是忽略 Σ 结构的 $g(Z_{is})$ 的传统估计量。因此,在这种情况下,偏误首项变成(见习题 19.7) 624

$$\frac{\kappa^2}{2}\sum_{r=1}^{q}h_r^2\left\{g_{rr}(z) + \Omega(z)^{-1}\sum_{t=1}^{T}\sum_{s=1}^{T}E[\sigma^{ts}g_{rr}(Z_{is}) \mid Z_{it}=z]\right\} \quad (19.104)$$

19.11 习 题

习题 19.1 对于单向误差成分误差结构,推导由式(19.3)定义的 $\hat{g}(x)$ 的渐近方差,即 $u_{it}=\mu_i+\nu_{it}$,其中 μ_i 是独立同分布 $(0,\sigma_\mu^2)$, ν_{it} 是独立同分布 $(0,\sigma_\nu^2)$,考虑如下的三种情况:

(i) $N\to\infty$, T 是有限的。

(ii) $T\to\infty$, N 是有限的。

(iii) N 和 T 都是无穷的。

习题 19.2 证实 v_N 的渐近方差确实是由式(19.17)给出的。

提示：

$$\begin{aligned}\operatorname{var}(v_N) &= \Omega(z)^{-2}N^{-1}\sum_t E[K^2_{h,Z_{it},z}c_t'\Sigma^{-1}E(u_iu_i' \mid z)\Sigma^{-1}c_t] + (\text{s.o.}) \\ &= \Omega(z)^{-2}N^{-1}\kappa^q(h_1,\cdots,h_q)^{-1}\sum_{t=1}^{T}f(z)E[\sigma^{tt} \mid Z_{it}=z] + (\text{s.o.}) \\ &= \kappa^2/[\Omega(z)Nh_1,\cdots,h_q] + (\text{s.o.})\end{aligned}$$

其中(s.o.)来自 $s \neq t$，且 $E[K_{h,z_{it},z}K_{h,z_{is},z}] = O(1)$，其小于 $E[K^2_{h,z_{it},z}] = O((h_1,\cdots,h_q)^{-1})$。

习题 19.3 推导式(19.42)。

提示：从式(19.40)开始，接着加上和减去 $\alpha_0(z) + [(Z_{it}-z)/h]'\alpha_1(z)$，然后结合这些项，且求解 $\begin{bmatrix}\hat{\alpha}_0(z)\\ \hat{\alpha}_1(z)\end{bmatrix} \equiv \begin{bmatrix}\hat{g}_{[l]}(z)\\ \hat{g}^{(1)}_{[l]}(z)\end{bmatrix}$。

625

习题 19.4 推导式(19.44)和式(19.45)。

习题 19.5 利用式(19.64)来证明式(19.65)。

习题 19.6 证明式(19.101)。

提示：首先，通过 Taylor 展开，我们有

$$A_{1N1} = N^{-1}\sum_{i=1}^{N}\sum_{t=1}^{T}K_{h,it,z}\sigma^{tt}(Z_{it}-z)'g^{(2)}(z)(Z_{it}-z) + R_N$$

(R_N 是 Taylor 展开中的余项)。那么，

$$E[A_{1N1}] = \kappa_2\sum_{t=1}^{T}\sigma^{tt}\sum_{r=1}^{q}h_r^2 g_{rr}(z) + o(\mid h \mid^2))$$

习题 19.7 证明式 (19.104)。

提示：来自 A_{1N1} 的项在习题 19.6 中被证明。来自 A_{2N} 的项通过 $\hat{g}_{[0]}(Z_{is}) - g(Z_{is})$ 的偏误首项得到。

第20章 非参数估计应用专题

在本章,我们将考虑文献中已经出现过的一些非参数方法的有趣应用。因为这些专题不太适合放在之前的任何一章,所以我们把它们结合在一起作为单独的一章,有关的方法在之前适当的章节都已提到。我们希望此处包括的这些应用能够加强非参数方法潜在的广泛应用,并且我们也鼓励读者从事自己的应用工作。 627

20.1 连续时间模型的非参数方法

20.1.1 连续时间模型的非参数估计

在本节,我们讨论连续时间扩散模型的非参数方法,我们关注扩散过程的非参数估计和参数扩散模型的非参数检验;参见 Cai 和 Hong(2003)以及 Fan(2005),他们提供了关于连续时间金融模型的非参数方法的综述。

连续时间模型在金融方面已经被广泛使用来捕捉重要经济变量的动态,如利率、汇率及股票价格。例如, 628 Black 和 Scholes(1973)提出的著名的"期权定价模型",以及 Cox, Ingersoll 和 Ross(1985)的"期限结构模型"。假定潜在的状态变量服从如下描述的一个扩散过程:

$$dX_t = \mu(X_t)\,dt + \sigma(X_t)\,dW_t \tag{20.1}$$

其中 $\mu(\cdot)$ 和 $\sigma^2(\cdot)$ 分别是 X_t 的"漂移"(瞬时均值)和"扩散"(瞬时方差),W_t 是一个标准的布朗运动。模

型化这些过程的一个参数方法是设定 $\mu(\cdot)$ 和 $\sigma^2(\cdot)$ 的参数函数形式，例如，$\mu(X_t,\theta)$ 和 $\sigma^2(X_t,\theta)$，其中 θ 是有限维的参数。在设定好 $\mu(\cdot)$ 和 $\sigma^2(\cdot)$ 的函数形式后，接着就可以着手估计 θ；参见 Gallant 和 Long(1997) 关于 θ 的最小距离估计以及 Ait-Sahalia(2002) 的近似的最大似然方法。然而，像所有参数方法一样，预先的设定可能与 DGP 是不一致的，因此，研究者可能倾向于非参数方法。

漂移和扩散项是 X_t 极小变化时($t\in[0,T]$)的前两个条件矩，即

$$\mu(X_t) = \lim_{\Delta\to 0} E\left[\frac{X_{t+\Delta}-X_t}{\Delta}\,\middle|\,X_t\right]$$

$$\sigma^2(X_t) = \lim_{\Delta\to 0} E\left[\frac{(X_{t+\Delta}-X_t)^2}{\Delta}\,\middle|\,X_t\right] \tag{20.2}$$

假定数据是等间隔的，$\{X_{\tau\Delta}\}_{\tau=1}^n$ 是可观测的离散样本。样本大小为 $n=T/\Delta$。使用离散数据，则可以分别通过 $\Delta^{-1}E[X_{(\tau+1)\Delta}-X_{\tau\Delta}\mid X_{\tau\Delta}]$ 和 $\Delta^{-1}E[(X_{(\tau+1)\Delta}-X_{\tau\Delta})^2\mid X_{\tau\Delta}]$ 来逼近 $\mu(x)$ 和 $\sigma^2(x)$。两种情况下的近似误差都是 $O(\Delta)$；详细情况参见 Stanton(1997) 以及 Fan 和 Zhang(2003)。上面的论证提示可以通过如下的式子估计 $\mu(\cdot)$ 和 $\sigma^2(\cdot)$：

$$\hat{\mu}(x) = \frac{1}{\Delta}\frac{\sum_{\tau=1}^{n}(X_{(\tau+1)\Delta}-X_{\tau\Delta})k_h(x-X_{\tau\Delta})}{\sum_{\tau=1}^{n}k_h(x-X_{\tau\Delta})} \tag{20.3}$$

$$\hat{\sigma}^2(x) = \frac{1}{\Delta}\frac{\sum_{\tau=1}^{n}(X_{(\tau+1)\Delta}-X_{\tau\Delta})^2k_h(x-X_{\tau\Delta})}{\sum_{\tau=1}^{n}k_h(x-X_{\tau\Delta})} \tag{20.4}$$

其中 $k_h(v)=h^{-1}k(v/h)$，$k(\cdot)$ 是满足式(1.10)具有有界支撑的二阶核。

定义

$$\hat{m}_1(x) = [\hat{\mu}(x)-\mu(x)]\hat{f}(x) \tag{20.5}$$

629 其中

$$\hat{f}(x) = \frac{1}{n\Delta}\sum_{t=1}^{n}k_h(X_{t\Delta}-x)$$

然后我们有

$$\hat{\mu}(x)-\mu(x) = \hat{m}_1(x)/\hat{f}(x)$$

使用我们在定理 1.4 中使用过的相同的证明，我们知道

$$\hat{f}(x) = f(x) + O(h^2+[\ln(n)/(nh)]^{1/2}) \quad \text{a.s.}$$

于是,我们有

$$\hat{\mu}(x)-\mu(x)=\hat{m}_1(x)/f(x)+o(\hat{m}_1(x)) \quad \text{a. s.}$$

假定$f(x)>0$,我们将证明

$$\mathrm{MSE}[\hat{m}_1(x)/f(x)]=O(\Delta^2+h^4+(nh\Delta)^{-1})$$

其暗示着

$$\hat{\mu}(x)-\mu(x)=O_p(\Delta+h^2+(nh\Delta)^{-1/2})$$

类似地,若我们定义

$$\hat{m}_2(x)=(\hat{\sigma}^2(x)-\sigma^2(x))\hat{f}(x) \tag{20.6}$$

则

$$\hat{\sigma}^2(x)-\sigma^2(x)=\hat{m}_2(x)/\hat{f}(x)=\hat{m}_2(x)/f(x)+o(\hat{m}_2(x)) \quad \text{a. s.}$$

我们也将证明

$$\mathrm{MSE}(\hat{m}_2(x))=O(\Delta^2+h^4+(nh\Delta)^{-1})$$

其暗示着$\hat{\sigma}^2(x)$是一致的。

Bandi 和 Phillips(2003)以及 Fan 和 Zhang(2003)建立了如下的结论。

定理 20.1 假定X_t是一个平稳β混合 Markov 过程,具有混合系数$\beta_\tau=\rho^\tau$,对某些$0<\rho<1$,边缘概率密度函数$f(x)$是有界且连续的,$\mu(x)$和$\sigma^2(x)$都是二次连续可微的,且$n\to\infty$,$T\to\infty$,$\Delta\to 0$,$h\to 0$ $\left(\text{回忆 } \kappa_2=\int v^2k(v)\mathrm{d}v,\ \kappa=\int k^2(v)dv\right)$,则($g^{(s)}(x)=\mathrm{d}^sg(x)/\mathrm{d}x^s, s=1,2$)

(i) $$E[\hat{m}_1(x)/f(x)]=\frac{h^2}{2}\kappa_2[\mu^{(2)}(x)+2\mu^{(1)}(x)f^{(1)}(x)/f(x)]+\frac{\Delta}{2}[\mu^{(1)}(x)\mu(x)+\mu^{(2)}(x)\sigma^2(x)]+o(\Delta+h^2)$$

且$\mathrm{var}[\hat{m}_1(x)/f(x)]=(nh\Delta)^{-1}\kappa\sigma^2(x)/f(x)[1+o(1)]$。

(ii) $$E[\hat{m}_2(x)/f(x)]=\frac{\Delta}{2}\left\{[2\mu(x)+(\sigma^2)^{(1)}(x)]\mu(x)+\left[2\mu^{(1)}(x)+\frac{1}{2}(\sigma^2)^{(2)}(x)\right]\sigma^2(x)\right\}+\frac{h^2}{2}\kappa_2[(\sigma^2)^{(2)}(x)+2\sigma^{(1)}(x)f^{(1)}(x)/f(x)]+o(\Delta+h^2)$$

且$\mathrm{var}[\hat{m}_2(x)/f(x)]=\frac{1}{nh}\kappa\sigma^4(x)/f(x)[1+o(1)]$。

630 定理20.1(i)的证明在20.1.5节被给出。(ii)的证明与(i)相似从而被省略。

注意定理20.1告诉我们$\hat{\mu}(x)$和$\hat{\sigma}^2(x)$有不同的收敛速度。$\hat{\sigma}^2(x)$的一致性仅要求当$n\to\infty$时(T可以被固定)$h\to 0$,而$\hat{\mu}(x)$的一致性要求一个更强的条件,即$T=n\Delta\to\infty$(这意味着$n\to\infty$,因为$\Delta\to 0$)。因此,若T不足够大,则$\mu(\cdot)$的准确估计比对$\sigma^2(\cdot)$的准确估计困难得多。

Bandi和Phillips(2003)建立了对于$\hat{\mu}(x)$和$\hat{\sigma}^2(x)$的几乎确定收敛以及渐近分布结果。在与定理20.1相同的条件下,并且假定$\Delta=o(h^2)$, Bandi和Phillips证明

$$\sqrt{Th}[\hat{\mu}(x)-\mu(x)-h^2\Gamma_{\mu}(x)]\to N\left(0,\frac{\kappa\sigma^2(x)}{f(x)}\right) \text{ 依分布收敛} \quad (20.7)$$

其中

$$\Gamma_{\mu}(x)=\kappa_2\left[\mu^{(1)}(x)f^{(1)}(x)/f(x)+\frac{1}{2}\mu^{(2)}(x)\right]$$

并且渐近方差可以通过$\kappa\hat{\sigma}^2(x)/\hat{f}(x)$被一致地估计。他们同样证明了

$$\sqrt{nh}[\hat{\sigma}^2(x)-\sigma^2(x)-h^2\Gamma_{\sigma^2}(x)]\to N\left(0,\frac{\kappa\sigma^4(x)}{f(x)}\right)\text{依分布收敛} \quad (20.8)$$

其中

$$\Gamma_{\sigma^2}(x)=\kappa_2\left[(\sigma^2)^{(1)}(x)f^{(1)}(x)/f(x)+\frac{1}{2}(\sigma^2)^{(2)}(x)\right]$$

并且渐近方差可以通过$\kappa\hat{\sigma}^4(x)/\hat{f}(x)$被一致地估计。注意当$\Delta=o(h^2)$时,与$\Delta$有关的偏误项是可以忽略的。因此,式(20.7)和式(20.8)有与h^2成比例的偏误首项。

式(20.7)和式(20.8)的证明在Bandi和Phillips(2003)以及Bandi和Moloche(2005)中被给出。推导如上结论使用的一个关键假定是X_t的平稳性。对于非平稳数据, Bandi和Phillips以及Moloche提出通过下式来估计$\mu(x)$和$\sigma^2(x)$:

$$\bar{\mu}(x)=\frac{\sum_{\tau=1}^{n}k_h(x-X_{\tau\Delta})\tilde{\mu}(X_{\tau\Delta})}{\sum_{\tau=1}^{n}k_h(x-Z_{\tau\Delta})}$$

$$\bar{\sigma}^2(x)=\frac{\sum_{\tau=1}^{n}k_h(x-X_{\tau\Delta})\tilde{\sigma}^2(X_{\tau\Delta})}{\sum_{\tau=1}^{n}k_h(x-Z_{\tau\Delta})} \quad (20.9)$$

其中 631

$$\tilde{\mu}(x) = \Delta^{-1}\sum_{\tau=1}^{n-1} w_{i,n}(x)(X_{(\tau+1)\Delta} - X_{\tau\Delta})$$

$$\tilde{\sigma}^2(x) = \Delta^{-1}\sum_{\tau=1}^{n-1} w_{i,n}(x)(X_{(\tau+1)\Delta} - X_{\tau\Delta})^2$$

有 $w_{i,n} = \mathbf{1}(|X_{\tau\Delta} - x| \leqslant b) \Big/ \sum_{\tau=1}^{n} \mathbf{1}(|X_{\tau\Delta} - x| \leqslant b)$，此处 $b = b_n$ 是一个依赖于 T 和 n 的平滑参数。他们同样建立了对于 $\bar{\mu}(x)$ 和 $\bar{\sigma}^2(x)$ 的一致和渐近结果。

对于非平稳数据，式(20.8)中的 $f(x)$ 必须被替换为所谓的局部时间，即

$$L_T(x) = \lim_{\Delta\to 0}\frac{1}{2\Delta}\int_0^1 \mathbf{1}(X_t \in A_\Delta)\mathrm{d}t$$

其中 $A_\Delta = (x-\Delta, x+\Delta)$，并且正态分布必须被替换为一个混合的正态分布。例如，$\sqrt{Th}[\hat{\mu}(x) - h^2\Gamma_\mu]$ 依分布收敛于一个混合正态随机变量，$\kappa^{1/2}[L_T(x)]^{-1}\mathcal{Z}$，其中 $\mathcal{Z}$ 是一个独立于局部时间 $L_T(x)$ 的标准正态随机变量。

式(20.1)给出的时间同质扩散模型的一个限制是它不允许时间效应。一个更加一般的时间依赖扩散模型被给定为

$$\mathrm{d}X_t = \mu(X_t, t) + \sigma(X_t, t)\mathrm{d}W_t \tag{20.10}$$

如同在式(20.2)中的情况，我们有

$$\mu(X_t, t) = \lim_{\Delta\to 0} E\left[\frac{X_{t+\Delta} - X_t}{\Delta}\,\middle|\, X_t\right]$$

$$\sigma^2(X_t, t) = \lim_{\Delta\to 0} E\left[\frac{(X_{t+\Delta} - X_t)^2}{\Delta}\,\middle|\, X_t\right] \tag{20.11}$$

其表明一个回归框架对于估计 $\mu(x,t)$ 和 $\sigma^2(x,t)$ 是适合的。

然而时间异质模型暗示着 X_t 是非平稳的，因此非参数方法的渐近结论将更难建立。在一般时间异质下非参数漂移和扩散估计量的渐近性质在现在是未知的，当然值得进一步研究。

相关的著作包括 Fan 和 Zhang(2003)，他们提出使用局部线性方法来估计下面的时间变化系数扩散模型。 632

$$\mathrm{d}X_t = [\alpha_0(t) + \alpha_1(t)X_t]\mathrm{d}t + \beta_0(t)\mathrm{d}W_t \tag{20.12}$$

其中 $\alpha_0(\cdot)$、$\alpha_1(\cdot)$ 和 $\beta_0(\cdot)$ 的函数形式是未设定的。

20.1.2 连续时间模型的非参数检验

利率、汇率和股票价格的动态经常被模型化为一个连续时间扩散过程。然

而,经济理论很少提供关于连续时间模型的函数形式。因此,对于参数连续时间模型的正确设定的检验是很重要的。Ait-Sahalia(1996)以及 Gallant 和 Tauchen(1996)发展了对于扩散模型的非参数和半参数检验,而 Fan, Zhang 和 Zhang(2001), Corradi 和 Swanson(2005), Hong 和 Li(2005),以及 Li 和 Tkacz(2006)提出了其他的检验。在接下来的两节,我们讨论 Ait-Sahalia 以及 Hong 和 Li 的检验。

20.1.3 Ait-Sahalia 的检验

Ait-Sahalia(1996)考虑如下连续时间平稳扩散过程:

$$dX_t = \mu_0(X_t) + \sigma_0(X_t)dW_t$$

其中 $\mu_0(X_t)$ 和 $\sigma_0(X_t)$ 是真实的漂移和扩散函数,W_t 是一个标准布朗运动。Ait-Sahalia 考虑了如下检验问题:对所有 x 和某些 $\theta_0 \in \Theta$,检验原假定 $\mu_0(x) = \mu(x,\theta_0)$ 以及 $\sigma_0(x) = \sigma^2(x,\theta_0)$,其中 Θ 是 $\mathbb{R}^q$ 的一个紧集。他建议通过检验一个从参数漂移和扩散函数 $\mu(x,\theta)$ 和 $\sigma^2(x,\theta)$ 推导出的参数边缘概率密度函数来间接地检验原假定。Ait-Sahalia 证明对应于(μ^2,σ^2)的边缘概率密度函数是

$$f(x,\theta) = \frac{\xi(\theta)}{\sigma^2(x,\theta)}\exp\left\{\int_{x_0}^{x}\frac{2\mu(v,\theta)}{\sigma^2(v,\theta)}dv\right\} \tag{20.13}$$

其中 x_0 是 X 支撑的一个内点,$\xi(\theta)$是一个正规化常数,以保证概率密度函数的积分为 1。令 $f_0(x)$表示真实但未知的边缘概率密度函数。可以通过第 1 章讨论过的核方法来估计 $f_0(x)$,我们把其表示为 $\hat{f}(\cdot)$,Ait-Sahalia 建议通过下式来估计 θ:

$$\hat{\theta} = \arg\min_{\theta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}[f(X_i,\theta) - \hat{f}(X_i)]^2$$

为了检验原假定,对几乎所有 x,$f(x,\theta) = f_0(x)$,检验统计量被给定为

$$\hat{M} = nh^{1/2}\left[\frac{1}{n}\sum_{i=1}^{n}(f(X_i,\hat{\theta}) - \hat{f}(X_i))^2\right] \tag{20.14}$$

在一些正则性条件下, Ait-Sahalia(1996)证明在原假定 $H_0: f(x,\theta) = f_0(x)$ 下,对几乎所有 x,有

$$[\hat{M} - \hat{E}_M]/\hat{V}_M^{1/2} \xrightarrow{d} N(0,1) \tag{20.15}$$

其中

$$\hat{E}_M = n^{-1}\left[\sum_{i=1}^{n}\hat{f}(X_i)\right]\int_{-\infty}^{\infty}k^2(v)dv$$

$$\hat{V}_M = 2\left[n^{-1}\sum_{i=1}^{n}\hat{f}(X_i)^2\right]\int\left[\int k(u)k(u+v)du\right]^2dv$$

633

20.1.4 Hong 和 Li 的检验

假定一个状态变量 X_t 服从如下形式的连续时间扩散过程：

$$dX_t = \mu_0(X_t, t) + \sigma_0(X_t, t)dW_t$$

其中 $\mu_0(X_t, t)$ 和 $\sigma_0(X_t, t)$ 是真实的漂移和扩散函数，W_t 是一个标准布朗运动。我们有兴趣检验原假定：μ_0 和 σ_0 都属于一些参数族，比如 $\mu_0 \in \mathcal{M}_u = \{\mu(\cdot, \cdot, \theta), \theta \in \Theta\}$，且 $\sigma_0 \in \mathcal{M}_\sigma = \{\sigma(\cdot, \cdot, \theta), \theta \in \Theta\}$，其中 Θ 是一个有界有限维的参数空间。则原假定被给定为对某些 $\theta_0 \in \Theta$，

$$H_0: P[\mu(X_t, t, \theta_0) = \mu_0(X_t, t), \sigma_0(X_t, t, \theta_0) = \sigma_0(X_t, t)] = 1 \tag{20.16}$$

备择假定是上面的概率严格小于1。基于（近似）参数模型来估计 θ_0 的方法参见 Gallant 和 Tauchen 以及 Ait-Sahalia(2002)。 634

我们将使用 $\{X_{\tau\Delta}\}_{\tau=1}^n$ 来检验上面的原假定，$\{X_{\tau\Delta}\}_{\tau=1}^n$ 是 $\{X_t\}$ 在时间跨度 T 间隔 Δ 上观测到的一个离散样本，所以样本大小是 $n = T/\Delta$。令 $p_0(x, t | y, s)$ 是 X_t 的转移密度，即给定 $X_s = y, s < t$ 时 $X_t = x$ 的条件概率密度函数。在 H_0 下，几乎处处有 $p_x(x, t | y, s, \theta_0) = p_{x,0}(x, t | y, s)$。因此，也可以通过比较 $p_x(x, t | y, s)$ 的参数和非参数估计量来检验 H_0。然而，Hong 和 Li(2005)提出的一种更简单的方法涉及引入下面的动态概率积分变换：

$$Z_\tau(\theta) = \int_{-\infty}^{X_\tau\Delta} p[x, \tau\Delta | X_{(\tau-1)\Delta}, (\tau-1)\Delta, \theta]dx, \quad \tau = 1, \cdots, n \tag{20.17}$$

根据 Rosenblatt(1952)，可以证明在 H_0 下，$\{Z_t = Z_t(\theta_0)\}_{t=1}^n$ 是独立同分布 $U[0,1]$（即均匀）。因此，可以通过检验 $\{Z_t\}$ 是否是独立同分布 $U[0,1]$ 序列来检验 H_0，用这种方法我们可以避免估计原模型，而原模型是很难得到的，因为封闭形式似然函数对于大部分参数扩散模型是不存在的。

Hong 和 Li(2005)建议对某些整数 j，检验 Z_t 与 Z_{t-j} 是否是独立同分布 $U[0,1]$ 的随机变量。令 $g_j(z_1, z_2)$ 表示 $\{Z_t, Z_{t-j}\}$ 的联合概率密度函数，则可以通过比较 $g_j(z_1, z_2)$ 与一个二维的均匀概率密度函数来检验 H_0，这个二维的均匀概率密度函数等于两个 $U[0,1]$ 的边缘概率密度函数的乘积，其中

$$\hat{g}(z_1, z_2) = \frac{1}{n-j}\sum_{t=j+1}^{n} k_h(z_1, \hat{Z}_\tau)k_h(z_2, \hat{Z}_{\tau-j}) \tag{20.18}$$

其中 $\hat{Z}_\tau = Z_\tau(\hat{\theta})$，且

$$k_h(x,y) = \begin{cases} h^{-1}k\left(\dfrac{y-x}{h}\right) \Big/ \displaystyle\int_{-x/h}^{1} k(v)\,\mathrm{d}v, & x \in [0,h) \\ h^{-1}k\left(\dfrac{y-x}{h}\right), & x \in [h, 1-h] \\ h^{-1}k\left(\dfrac{y-x}{h}\right) \Big/ \displaystyle\int_{-1}^{(1-x)/h} k(v)\,\mathrm{d}v, & x \in (1-h, 1] \end{cases} \quad (20.19)$$

注意 $k_h(x,y)$ 是一个边界修正核函数，$k(\cdot)$ 是满足式(1.10)、具有支撑 [−1,1]的一个二阶核(参见式(1.43)关于不需要在支撑[−1,1]上的一个普通的边界修正核)。Hong 和 Li(2005)建议基于下式来检验 H_0

635

$$\begin{aligned} \hat{M}(j) &= \int_0^1 \int_0^1 [\hat{g}(z_1, z_2) - 1]^2 \mathrm{d}z_1 \mathrm{d}z_2 \\ &= \frac{1}{(n-j)^2} \sum_{t=j+1}^{n} \sum_{s=j+1}^{n} \int_0^1 \int_0^1 [k_h(z_1, \hat{Z}_t) k_h(z_2, \hat{Z}_{t-j}) - 1] \\ &\quad \times [k_h(z_1, \hat{Z}_s) k_h(z_2, \hat{Z}_{s-j}) - 1] \mathrm{d}z_1 \mathrm{d}z_2 \end{aligned}$$

下面我们建议通过在 $\hat{M}(j)$ 中移除 $t=s$ 的项来得到 $\hat{M}(j)$ 的一个轻微的修改，即

$$\begin{aligned} \tilde{M}(j) &= \frac{1}{(n-j)^2} \sum_{t=j+1}^{n} \sum_{s \neq t, s=j+1}^{n} \int_0^1 \int_0^1 [k_h(z_1, \hat{Z}_t) k_h(z_2, \hat{Z}_{t-j}) - 1] \\ &\quad \times [k_h(z_1, \hat{Z}_s) k_h(z_2, \hat{Z}_{s-j}) - 1] \mathrm{d}z_1 \mathrm{d}z_2 \end{aligned} \quad (20.20)$$

Hong 和 Li(2005)证明了接下来的结论。在 H_0 下，

$$nh\tilde{M}(j)/\hat{\sigma}_{HL} \to N(0,1) \text{ 依分布收敛} \quad (20.21)$$

其中

$$\hat{\sigma}_{HL}^2 = 2\left[\int_{-1}^{1} \left[\int_{-1}^{1y} [k(u+v)k(v)\mathrm{d}v\right]^2 \mathrm{d}u\right]^2$$

式(20.21)的证明在 Hong 和 Li(2005)的定理 1 的证明中给出。我们修改的检验与 Hong 和 Li 的原始检验 $\hat{M}(j)$ 的唯一差别是我们修改的检验 $\tilde{M}(j)$ 没有一个非零的中心项。

Florens-Zmirou(1993)首先给出了扩散 $\sigma^2(\cdot)$ 的非参数估计的渐近理论；但是，她没有考虑估计漂移项 $\mu(\cdot)$ 的问题。Jiang 和 Knight(1997)考虑了 $\sigma^2(\cdot)$ 和 $\mu(\cdot)$ 的非参数估计。他们获得了 $\hat{\mu}(x)$ 的一致估计，但是没有提供收敛速度或者是 $\hat{\mu}(x)$ 的渐近分布。Fan 和 Yao(1998)考虑使用局部线性核回归来估计 $\sigma^2(\cdot)$。Ait-Sahalia(1996)提出在漂移函数的参数设定下估计扩散过程的一个半参数过程。Stanton(1997)提出了非参数估计 $\sigma^2(\cdot)$ 和 $\mu(\cdot)$ 的

高阶近似，而 Fan 和 Zhang(2003)对于高阶近似进行了更深入的分析，并且推导了非参数扩散和漂移估计量的收敛速度。Bandi 和 Nguyen(1999)认为对于扩散和漂移的任何阶数的近似表现出同样的收敛速度，并且有相同的渐近方差，所以和计算问题相连接的渐近理论建议在实际中可以简单利用一阶近似。Bandi 和 Phillips(2003)首先给出了在一般条件下非参数漂移估计量的渐近分布理论(例如，允许非平稳)。Bandi 和 Moloche(2005)将 Bandi 和 Phillips 的工作扩展到多元扩散模型。一般时间异质扩散过程的非参数估计的渐近理论仍然是未知的。 636

因为篇幅的限制，我们不讨论跳跃扩散模型，虽然它对于模型化出现在金融数据中的大的冲击和波动性来说是有用的；关于跳跃扩散模型参见 Cai 和 Hong(2003)和其中的参考文献。

20.1.5 证明

在本节，我们概述定理20.1(i)的证明。

式20.1(i)的证明：由式(20.1)，我们有

$$X_{(t+1)\Delta}-X_{t\Delta}=\int_{t\Delta}^{(t+1)\Delta}\mu(X_u)\,\mathrm{d}u+\int_{t\Delta}^{(t+1)\Delta}\sigma(X_u)\,\mathrm{d}W_u$$

将上式代入 $\hat{g}(x)\overset{\mathrm{def}}{=}\hat{\mu}(x)\hat{f}(x)$，根据平稳性以及以下事实：

$$E\left[\int_{t\Delta}^{(t+1)\Delta}\sigma(X_u)\,\mathrm{d}W_u \mid X_{t\Delta}\right]=0$$

(因为 $u\geqslant t\Delta$)，我们有

$$\begin{aligned}E[\hat{g}(x)]&=\frac{1}{\Delta}E\left[k_h(X_{t\Delta}-x)\int_{t\Delta}^{(t+1)\Delta}\mu(X_u)\,\mathrm{d}u\right]\\&=\frac{1}{\Delta}E\left\{k_h(X_{t\Delta}-x)\int_{t\Delta}^{(t+1)\Delta}[\mu(X_u)-\mu(X_{t\Delta})]\,\mathrm{d}u\right\}\\&\quad+E[k_h(X_{t\Delta}-x)\mu(X_{t\Delta})]\\&=A_{1n}+A_{2n}\end{aligned}\tag{20.22}$$

其中 A_{1n} 和 A_{2n} 的定义应该是显而易见的。 637

使用 Ito 引理以及增加和减去一些项(见附录 A 中的式(A.14))，我们有

$$
\begin{aligned}
\mu(X_u)-\mu(X_{t\Delta}) &= \int_{t\Delta}^{u}\mu^{(1)}(X_s)\mu(X_s)\,\mathrm{d}s \\
&\quad+\frac{1}{2}\int_{t\Delta}^{u}\mu^{(2)}(X_s)\sigma^2(X_s)\,\mathrm{d}s+B_n \\
&=\mu^{(1)}(X_{t\Delta})\mu(X_{t\Delta})(u-t\Delta) \\
&\quad+(1/2)\mu^{(2)}(X_{t\Delta})\sigma^2(X_{t\Delta})(u-t\Delta) \\
&\quad+\int_{t\Delta}^{u}[\mu^{(1)}(X_s)\mu(X_s)-\mu^{(1)}(X_{t\Delta})\mu(X_{t\Delta})]\,\mathrm{d}s \\
&\quad+(1/2)\int_{t\Delta}^{u}[\mu^{(2)}(X_s)\sigma^2(X_s)-\mu^{(2)}(X_{t\Delta})\sigma^2(X_{t\Delta})]\,\mathrm{d}s+B_n
\end{aligned}
\tag{20.23}
$$

其中
$$B_n=\int_{t\Delta}^{u}\mu^{(1)}(X_u)\sigma(X_u)\,\mathrm{d}W_u$$

对式(20.23)进行积分得到(参见习题20.1)

$$
\begin{aligned}
&\int_{t\Delta}^{(t+1)\Delta}[\mu(X_u)-\mu(X_{t\Delta})]\,\mathrm{d}u \\
&\quad=\frac{\Delta^2}{2}\mu^{(1)}(X_{t\Delta})\mu(X_{t\Delta})+\frac{\Delta^2}{4}\mu^{(2)}(X_{t\Delta})\sigma^2(X_{t\Delta}) \\
&\quad\quad+\int_{t\Delta}^{(t+1)\Delta}\left[\int_{t\Delta}^{u}[\mu^{(1)}(X_s)\mu(X_s)-\mu^{(1)}(X_{t\Delta})\mu(X_{t\Delta})]\,\mathrm{d}s\right]\mathrm{d}u \\
&\quad\quad+\frac{1}{2}\left[\int_{t\Delta}^{(t+1)\Delta}\int_{t\Delta}^{u}[\mu^{(2)}(X_s)\sigma^2(X_s)-\mu^{(2)}(X_{t\Delta})\sigma^2(X_{t\Delta})]\,\mathrm{d}s\right]\mathrm{d}u \\
&\quad\quad+\int_{t\Delta}^{(t+1)\Delta}B_n\,\mathrm{d}u
\end{aligned}
\tag{20.24}
$$

将式(20.24)代回 A_{1n},且注意 $E(B_n)=E[E(B_n|X_{t\Delta})]=0$,我们得到

$$A_{1n}=\frac{\Delta}{2}E\left\{k_h(X_{t\Delta},x)\left[\mu^{(1)}(X_{t\Delta})\mu(X_{t\Delta})+\frac{1}{2}\mu^{(2)}(X_{t\Delta})\sigma^2(X_{t\Delta})\right]\right\}+O(\Delta^2)$$

638 其中 $O(\Delta^2)$ 来自双重积分项,因为两个积分范围都是 $O(\Delta)$。则根据一个变量变换定理,我们有

$$
\begin{aligned}
&E\left\{k_h(X_{t\Delta}-x)\left[\mu^{(1)}(X_{t\Delta})\mu(X_{t\Delta})+\frac{1}{2}\mu^{(2)}(X_{t\Delta})\sigma^2(X_{t\Delta})\right]\right\} \\
&\quad=\int f(x+hv)k(v)\left[\mu^{(1)}(x+hv)\mu(x+hv)\right. \\
&\quad\quad\left.+\frac{1}{2}\mu^{(2)}(x+hv)\sigma^2(x+hv)\right]\mathrm{d}v \\
&\quad=f(x)\left[\mu^{(1)}(x)\mu(x)+\frac{1}{2}\mu^{(2)}(x)\sigma^2(x)\right]+O(h^2)
\end{aligned}
$$

因此，我们得到

$$A_{1n}=\frac{\Delta}{2}f(x)\left[\mu^{(1)}(x)\mu(x)+\frac{1}{2}\mu^{(2)}(x)\sigma^2(x)\right]+o(\Delta)$$

类似地，我们有

$$\begin{aligned}\Delta A_{2n}&=\int k_h(X_{t\Delta}-x)\mu(X_{t\Delta})f(X_{t\Delta})\mathrm{d}X_{t\Delta}\\&=\int k(v)\mu(x+hv)f(x+hv)\mathrm{d}v\\&=\mu(x)f(x)+\frac{h^2}{2}\left[\int k(v)v^2\mathrm{d}v\right]\\&\quad\times[\mu^{(2)}(x)f(x)+2\mu^{(1)}(x)f^{(1)}(x)+\mu(x)f^{(2)}(x)]+o(h^2)\end{aligned}$$

总结以上，我们已经证明了

$$\begin{aligned}E\left[\frac{\hat{m}_1(x)}{f(x)}\right]&=f(x)^{-1}\{A_{1n}+A_{2n}-\mu(x)E[\hat{f}(x)]\}\\&=\frac{\Delta}{2}[\mu^{(1)}(x)\mu(x)+(1/2)\mu^{(2)}(x)\sigma^2(x)]\\&\quad+\frac{h^2}{2}\kappa_2[\mu^{(2)}(x)+2\mu^{(1)}(x)f^{(1)}(x)/f(x)]+o(h^2+\Delta)\end{aligned}$$

接下来，令

$$Z_{nt}=k_h(X_{t\Delta}-x)\int_{t\Delta}^{(t+1)\Delta}\sigma(X_u)\mathrm{d}W_u$$

可以证明，$\mathrm{var}(\hat{m}_1(x))$的首项来自 $\mathrm{var}(\hat{m}_3(x))$（见习题 20.2），其中

$$\hat{m}_3(x)=(n\Delta)^{-1}\sum_{t=1}^{n}Z_{nt}$$

现在，

$$\mathrm{var}(\hat{m}_3(x))=(n\Delta)^{-2}\left\{\sum_{t=1}^{n}\mathrm{var}(Z_{nt})+0\right\}=(n\Delta^2)^{-1}\mathrm{var}(Z_{nt})$$

（上式成立是因为 $\mathrm{cov}(Z_{nt},Z_{ns})=0,t\neq s$）。因此，我们有

$$\begin{aligned}\mathrm{var}(\hat{m}_3(x))&=(n\Delta^2)^{-1}\mathrm{var}\left\{\int_{t\Delta}^{(t+1)\Delta}k_h(X_{t\Delta}-x)\sigma(X_u)\mathrm{d}W_u\right\}\\&=(n\Delta^2)^{-1}\left\{E\left[\left(\int_{t\Delta}^{(t+1)\Delta}k_h(X_{t\Delta}-x)\sigma(X_u)\mathrm{d}W_u\right)^2\right]+(\mathrm{s.o.})\right\}\\&=(n\Delta^2)^{-1}\left\{E\left[\int_{t\Delta}^{(t+1)\Delta}k_h^2(X_{t\Delta}-x)\sigma^2(X_u)\mathrm{d}u\right]+(\mathrm{s.o.})\right\}\\&=(n\Delta^2)^{-1}\left\{\int f(x_{t\Delta})k_h^2(x_{t\Delta}-x)\left[\int_{t\Delta}^{(t+1)\Delta}\sigma^2(x_{t\Delta})\mathrm{d}u\right]\mathrm{d}x_{t\Delta}+(\mathrm{s.o.})\right\}\end{aligned}$$

639

$$= (n\Delta^2)^{-1}\left\{\left[h^{-1}f(x)\sigma^2(x)\int k^2(v)dv\right]\Delta + (\text{s.o.})\right\}$$

$$= (nh\Delta)^{-1}\{\kappa f(x)\sigma^2(x) + o(1)\}$$

因此，

$$\text{var}(\hat{m}_3(x)/f(x)) = (nh\Delta)^{-1}\kappa\sigma^2(x)/f(x)[1+o(1)]$$

这就完成了定理 20.1(i)的证明。 □

20.2 平均处置效应的非参数估计

平均处置效应的测量，最初限定在医学药剂反应关系的评价中，今天它已被广泛应用于一系列学科。评价由于战争导致的人力资本损失(Ichino 和 Winter-Ebmer(1998))以及评价职业培训项目的效率(Lechner(1999), Black, Smith, Berger 和 Noel(2003))是其中的两个例子。

640 一个测量处置效应的流行方法包括一个"倾向得分"的估计。倾向得分的估计(即接受处理的条件概率)最初是使用参数指数模型，如 Logit 或者 Probit 模型来进行的。近期，Hahn(1998)，Hirano，Imbens 和 Ridder(2003)提出平均处置效应的非参数估计，两种方法都是基于序列方法(见第 15 章的介绍)。虽然序列方法已经能够通过利用示性函数方法来很好地处理离散协变量，但很难将例如那些在第 4 章中介绍的平滑离散变量的技术扩展到非参数序列框架。给定被用来评估处置效应的数据集经常包含分类数据的优势，Li，Racine 和 Wooldridge(2005)建议通过核方法来估计处置效应、平滑离散和连续协变量，并且使用数据驱动的交错鉴定法来选择平滑参数；下面我们概述这种处理。正如在第 4 和第 5 章中讨论的，这种方法的一个限定性的特征是它允许平滑掉不相关的协变量，它们是连续还是离散，导致了在有限样本背景中的效率增益。

有关涉及倾向得分估计的相关文献，参见 Heckman，Ichimura 和 Todd (1997,1998)，以及 Dehejia 和 Wahba(1999)。也可参见 Vytlacil 和 Heckman (2005)，他们提出使用边际处置效应来统一处置效应的非参数文献以及使用一个政策不变参数的非参数模拟来进行结构估计的计量经济学文献。虽然此处我们没有强调潜在内生性，我们仍要提醒感兴趣的读者参见 Abadie(2003)，他考虑了处理响应模型的半参数工具变量估计。

20.2.1 模型

我们使用一个哑变量，$t_i \in \{0,1\}$，来表示一个个体是否接受处理，即 $t_i=1$ 是接受处理，而 $t_i=0$ 是没有接受处理。令 $Y_i(t_i)$ 表示结果，则

$$Y_i = t_i Y_i(1) + (1 - t_i) Y_i(0), \quad i = 1, \cdots, n$$

人们关心的是平均处置效应,它被定义为

$$\tau = E[Y_i(1) - Y_i(0)]$$

令 X_i 表示一个预处理变量向量。对于每个个体 i,我们观察到 $Y_i(0)$ 或者 $Y_i(1)$,但不是两者都观测到。因此,在没有其他假定的情况下,处置效应不是一致可估计的。一个流行的假定是"非混杂条件"(unconfoundedness condition)(Rosenbaum 和 Rubin(1983)),其阐述了

641

假定 20.1 条件于 X_i,处理指标 t_i 独立于潜在的结果。

存在协变量时,我们通过 $\tau(x) = E[Y_i(1) - Y_i(0) | X = x]$ 来定义条件平均处置效应。在假定 20.1 的条件下,我们可以容易地证明

$$\tau(x) = E[Y_i | t_i = 1, X_i = x] - E[Y_i | t_i = 0, X_i = x] \quad (20.25)$$

式(20.25)右边的两项可以通过任一非参数估计技术被一致估计。因此,存在协变量时平均处置效应可以通过简单平均 $\tau(x)$ 而获得,其被给定为

$$\tau = E[\tau(X_i)]$$

令 $E(Y_i | X_i, t_i)$ 被表示为 $g(X_i, t_i)$,则我们有

$$Y_i = g(X_i, t_i) + u_i \quad (20.26)$$

其中 $E(u_i | X_i, t_i) = 0$。

定义 $g_0(X_i) = g(X_i, t_i = 0)$ 和 $g_1(X_i) = g(X_i, t_i = 1)$,我们可以重写式(20.26)为

$$\begin{aligned} Y_i &= g_0(X_i) + [g_1(X_i) - g_0(X_i)] t_i + u_i \\ &= g_0(X_i) + \tau(X_i) t_i + u_i \end{aligned} \quad (20.27)$$

其中 $\tau(X_i) = g_1(X_i) - g_0(X_i)$。

由式(20.27)我们可以证明

$$\tau(X_i) = \operatorname{cov}(Y_i, t_i | X_i) / \operatorname{var}(t_i | X_i)$$

令
$$\mu_i \equiv \mu(X_i) = P(t_i = 1 | X_i) \equiv E(t_i | X_i)$$

(因为 $t_i = \{0, 1\}$),我们可以写为

$$\tau = E[\tau(X_i)] = E\left\{\frac{(t_i - \mu_i) Y_i}{\operatorname{var}(t_i | X_i)}\right\} \quad (20.28)$$

因为 $\mu(X_i) = P(t_i = 1 | X_i) \equiv E(t_i | X_i)$,可以使用一个条件概率估计量或者是一个条件均值估计量来估计 $\mu(X_i)$。在本章我们使用后者。令 $\hat{t}(X_i)$ 是 μ_i 的非参数估计量,定义为

$$\hat{t}(X_i) = \frac{\sum_{j=1}^{n} t_j K_{n,ij}}{\sum_{j=1}^{n} K_{n,ij}} \quad (20.29)$$

642 其中 $K_{n,ij}$ 是对于所有条件变量的一个乘积核函数。通过注意有 $\mathrm{var}(t_i \mid X_i) = \mu_i(1-\mu_i)$，平均处置效应可以通过下式来估计：

$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}\frac{(t_i - \hat{t}(X_i))Y_i M_{ni}}{\hat{t}(X_i)(1-\hat{t}(X_i))} \equiv \frac{1}{n}\sum_{i=1}^{n}\left[\frac{t_i Y_i}{\hat{t}(X_i)} - \frac{(1-t_i)Y_i}{1-\hat{t}(X_i)}\right]M_{ni} \tag{20.30}$$

其中 $M_{ni} = M_n(X_i)$ 是一个修剪集，可以修剪掉边界附近的观测值。

除了有修剪函数 M_{ni} 的存在以及使用核方法而不是序列法，式(20.30)与 Hirano 等(2003)考虑的估计量是完全相同的。Li 等(2005)证明 $\hat{\tau}$ 与 Hahn (1998)以及 Hirano 等提出的基于序列方法的估计量有相同的渐近分布。现在我们转向一个实证的应用。

20.2.2 一个应用：评估右心导管检查的效率

通过右心导管检查(RHC)来对心脏功能进行直接的测量被用于对一些病危病人的诊断和引导适当治疗。传统观点认为 RHC 的使用导致了更好的结果。

Connors 等(1996)报告了一项研究，其被设计用来评估 RHC 对 5 735 名需要被特别护理的病危病人的效率。他们进行了一个前瞻性群组研究，检验在前 24 小时住院治疗使用 RHC 与随后生存之间的关系。他们总结出与传统观点相反的结论，在那些接受 RHC 的人中，死亡的可能性更高，即使在认真控制了许多风险因素后。

在接下来的研究中，Lin，Psaty 和 Kronmal(1998)进行了进一步的分析来决定这种结论是否可能是由于未测量的混杂因素的存在，并且他们得出结论，这种情况不可能，虽然他们也总结出死亡的风险可能没有 Connors 等(1996)发现的那么高。

然而这些研究的结论依赖于其中使用的参数模型的适当性。Li 等(2005)
643 调查了这种未预期到的结果事实上是否反映了所实施的限制性的参数设定。

为了接下来的分析，Li 等(2005)使用了如下的变量：

(i) Y：结果——若 180 天内死亡发生则为 1，其他为 0。

(ii) T：治疗——若病人住院时接受了一个 Swan-Ganz 导管则治疗则为 1，其他为 0。

(iii) X_1：性别——女性为 0，男性为 1。

(iv) X_2：种族——黑人为 0，白人为 1，其他为 2。

(v) X_3：收入——11 000 以下为 0，11 000—25 000 为 1，25 000—50 000 为 2，超过 50 000 为 3。

(vi) X_4:初级疾病分类——急性呼吸衰竭为1,充血性心力衰竭为2,慢性阻塞性肺疾病为3,肝硬化为4,结肠癌为5,昏迷为6,肺癌为7,恶性多器官功能衰竭为8,脓毒症多器官功能衰竭为9。

(vii) X_5:二级疾病分类——肝硬化为1,结肠癌为2,昏迷为3,肺癌为4,恶性多器官功能衰竭为5,脓毒症多器官功能衰竭为6,NA为7。

(viii) X_6:医疗保险——医疗救助为1,医疗保险为2,医疗救助和医疗保险为3,没有保险为4,私人为5,私人医疗保险为6。

(ix) X_7:年龄——年龄(从精确到两位小数存储的年/月/日的数据形式转化为年)。

我们观察到这些变量大多数被自然地分类,因此这个应用很适合在第4章中介绍的混合数据核方法。

基于估计量的参数倾向得分给出的平均处置效应为0.072,式(20.30)给出的非参数倾向得分得到的平均处置效应为-0.001。Li等(2005)也计算了参数和非参数倾向得分的混淆矩阵,并且发现非参数方法对于是否接受RHC处理正确地预测了69.9%,而参数方法正确地预测了66.7%。对于这个数据集,即使可能的单元数目为18 144,其远超过记录的数量($n=5\,735$),非参数方法也在预测谁接受了治疗而谁没有接受方面好于参数Logit模型。参数方法正确预测了5 735个病人中的3 828个,而非参数方法正确地预测了3 976个,多正确预测了148个病人。加权估计的参数和非参数版本之间的差别反映了正确分类病人的增加数目以及在估计出的接受治疗概率上的差别。参数模型给出的上升的风险从接受RHC而增加的7%下降到使用非参数方法时的大约0%。

644

为了评价这些差别是否是由于偶然导致的,Li等(2005)使用了如下的再抽样过程来构建$\hat{\tau}$的样本分布。令$z=\{y,x\}$,

(i) 有放回地从$\{Z_j\}_{j=1}^n$中随机抽取Z_i^*,把$\{Z_i^*\}_{i=1}^n$称为自举样本。

(ii) 使用自举样本来计算自举统计量$\hat{\tau}^*$,与在获得$\hat{\tau}$时使用的是同样的交错鉴定平滑参数。

(iii) 重复(i)和(ii)多次,如$B=399$次。$\{\hat{\tau}_j^*\}_{j=1}^B$的经验累积分布函数可以被用来逼近$\hat{\tau}$的有限样本分布。

Li等(2005)获得了对于参数方法自举95%的收敛误差边界为[0.044,0.099],非参数方法是[-0.038,0.011]。从而,他们推翻了非预期的参数结果且得出与传统观点一致的结论,病人接受RHC治疗不会遭遇死亡率的增加。为了进一步检查这个结果,Li等(2005)也进行了一些敏感性分析。使用似然交错鉴定而不是最小二乘交错鉴定得到95%的收敛误差边界为[-0.034,

0.013]。出于关心非参数结果可能反映“过度拟合”,他们计算了去一核预测,
645 并再次计算自举误差边界。对于最小二乘交错鉴定去一估计他们获得的95%的收敛误差边界为[-0.015,0.037],而对于似然交错鉴定去一估计他们获得的95%的收敛误差边界为[-0.007,0.039]。

这些误差边界表明参数模型显示接受 RHC 导致死亡风险的显著增加,而非参数模型得到的结果是差异不显著。这并非反映了由于使用非参数而不是参数倾向得分所导致的任何效率损失,这可以从混淆矩阵的样本外预测结果的比较中看出,因为一个正确设定的参数模型可以被预期表现优于非参数模型。

20.3 拍卖模型的非参数估计

20.3.1 一级价格拍卖模型的估计

自 Paarsch(1992)创造性的工作以来,计量经济学家在估计拍卖模型时已经完全采用结构计量经济学方法。早期的工作主要基于参数模型;参见 Donald 和 Paarsch(1993,1996),以及 Laffont, Ossard 和 Vuong(1995)。第一个我们知道的非参数方法是 Guerre, Perrigne 和 Vuong(2000)的著作。我们推荐感兴趣的读者参见 Paarsch 和 Hong(2006)对拍卖数据结构计量经济学的一个详尽的处理。

假定一个单一不可分的物体被拍卖,所有的出价是同时收集的。标的被出售给最高出价人,给定这个出价至少和一个保留价格 p_0 一样高。在独立私人价值(IPV)范式中,有 $i=1,\cdots,I$ 个投标者。个体 i 不知道其他人的私人价值 $v_j(j\neq i)$;然而,所有投标者的私人价值被假定是从一个具有绝对连续概率密度函数 $f(\cdot)$ 的已知累积分布函数 $F(\cdot)$ 中随机抽取的,绝对连续概率密度函数 $f(\cdot)$ 有紧支撑 $[\underline{v},\bar{v}]\subset\mathbb{R}$,$p_0\in[\underline{v},\bar{v}]$。

不完全信息下唯一对称微分贝叶斯纳什均衡被 Riley 和 Samuelson(1981)特征化为如下简洁的形式。第 i 个投标者的均衡标价 b_i 被给出为

$$b_i=v_i-\frac{1}{(F(v_i))^{I-1}}\int_{p_0}^{v_i}(F(u))^{I-1}\mathrm{d}u\equiv s(v_i,F,I,p_0) \qquad (20.31)$$

646 给定 $v_i\geqslant p_0$。这种策略是通过求解 $s(\cdot)$ 中的一阶微分方程而得到的,给定为

$$1=(v_i-s(v_i))(I-1)\frac{f(v_i)}{F(v_i)}\frac{1}{s^{(1)}(v_i)} \qquad (20.32)$$

其具有边界条件 $s(p_0)=p_0$。均衡策略(20.31)在 $[p_0,v]$ 上关于 v_i 是严格递增的。

在实践中，出价是可以被观测的，而私人价值是不可观测的。假定保留价格 p_0 不是粘性的，即 $p_0 = \underline{v}$，有 $s(v_{\underline{v}}) = \underline{v}$。令 $G(\cdot)$ 表示 b_i 的累积分布函数，$g(\cdot)$ 表示它的概率密度函数。对每一个 $b \in [\underline{b}, \bar{b}] = [\underline{v}, s(\bar{v})]$，因为 $b = s(v)$，所以我们有

$$G(b) = P[\tilde{b} \leqslant b] = P[\tilde{v} \leqslant s^{-1}(b)] = F(s^{-1}(b)) = F(v)$$

$G(\cdot)$ 是连续的并且 $g(b) = f(v)/s^{(1)}(v)$，其中 $v = s^{-1}(b)$。取比值为

$$g(b)/G(b) = f(v)/[F(v)s^{(1)}(v)]$$

式(20.32)可以变成

$$v_i = b_i + \frac{1}{I-1}\frac{G(b_i)}{g(b_i)} \equiv \xi(b_i, G, I) \tag{20.33}$$

方程(20.33)证明个人私人价值 v_i 可以被表示成个人均衡出价 b_i 的一个函数。因此，如果可以获得 $G(\cdot)$ 和 $g(\cdot)$ 的一致估计，就可以通过式(20.33)由 b_i 估计出不可观测的私人价值 v_i。Guerre 等(2000)进一步证明 $G(\cdot)$ 是唯一可识别的，给定 $G(b_1, \cdots, b_I) = \prod_{i=1}^{I} G(b_i)$，$\xi(\cdot, G, I)$ 在 $[\underline{b}, \bar{b}]$ 上是严格递增的，且它的逆是可微的。Guerre 等(2000)建议采用如下的两阶段估计过程。假定有 L 个同质拍卖和相同数目的投标者 I。在第一步先通过下面的式子估计 $G(\cdot)$ 和 $g(\cdot)$：

$$\tilde{G}(b) = \frac{1}{IL}\sum_{l=1}^{L}\sum_{i=1}^{I}\mathbf{1}(B_{il} \leqslant b)$$

$$\tilde{g}(b) = \frac{1}{ILh}\sum_{l=1}^{L}\sum_{i=1}^{I}k\left(\frac{B_{il} - b}{h}\right) \tag{20.34}$$

其中 $k(\cdot)$ 是具有紧支撑的一个核函数，h 是平滑参数。

为了避免在边界处的估计偏差，Guerre 等(2000)建议使用由下式定义的修剪数据：

$$\hat{V}_{il} = B_{il} + \tilde{G}(B_{il})/[(I-1)\tilde{g}(B_{il})]$$

如果

$$B_{\min} + \rho/h/2 \leqslant B_{il} \leqslant B_{\max} - \rho h/2 \tag{20.35}$$

其中 ρ 是支撑 $k(\cdot)$ 的(有限)长度，$B_{\min}$ 和 $B_{\max}$ 是最小出价和最大出价。则可以通过下式估计 $f(\cdot)$：

647

$$\hat{f}(v) = \frac{1}{ILh}\sum_{l=1}^{L}\sum_{i=1}^{I}k\left(\frac{\hat{V}_{il} - v}{h_f}\right) \tag{20.36}$$

其中 $\hat{V}_{il}$ 在式(20.35)中被定义，其剔除了边界观测值。

假定 $k(\cdot)$ 是一个 ν 阶核，$f(\cdot)$ 有 ν 个有界连续导数，我们知道估计 $f(\cdot)$ 的最优一致收敛速率是 $(n/\ln n)^{\nu/(2\nu+1)}$，其中 $n=IL$（见 Stone(1982)）。然而，在这种情况下，b_i 是可观测的，而 v_i 是不可观测的。Guerre 等(2000)证明在这种情形下，$\hat{f}(\cdot)$ 对 $f(\cdot)$ 的最优一致收敛速率是 $(n/\ln n)^{\nu/(2\nu+3)}$。利用 $h=c(n/\ln n)^{1/(2\nu+3)}$ 及 $h_f=c_f(n/\ln n)^{1/(2\nu+3)}$，以上估计量 $\hat{f}(\cdot)$ 达到了这个最优的速率。这个收敛速率改变的发生是因为一个生成回归元的出现。根据一个 Lipschitz 类型条件，收敛速率与当使用核方法估计导函数时达到的是一样的。

Guerre 等(2000)将他们的结论也扩展到了异质拍卖的情况以及在不同的拍卖中投标者可能改变的情况。考虑 L 个异质拍卖。令 X_l 表示第 l 个被拍卖标的的相关特征向量，I_l 表示在第 l 个拍卖上的投标者数目。现在，对于第 l 个拍卖，投标者的私人价值 V_{il} 的分布是私人价值给定为 (X_l,I_l) 下的条件累积分布函数 $F(\cdot\mid X_l,I_l)$。类似地，$G(\cdot)$ 变成 $G(\cdot\mid X_l,I_l)$，且式(20.31)和式(20.33)变为

$$
\begin{aligned}
B_{il} &= V_{ii} - \frac{1}{(F(V_{il}\mid X_l,I_l))^{I-1}}\int_{p_0}^{V_{ij}}(F(u\mid X_l,I_l))^{I-1}\mathrm{d}u \\
&\equiv s(V_{il},X_l,I_l) \qquad (20.37)
\end{aligned}
$$

$$
\begin{aligned}
V_{il} &= B_{il} + \frac{1}{I_l-1}\frac{G(B_{il}\mid X_l,I_l)}{g(B_{il}\mid X_l,I_l)} = B_{il} + \frac{1}{I_l-1}\frac{G(B_{il},X_l,I_l)}{g(B_{il},X_l,I_l)} \\
&\equiv \xi(B_{il},X_l,I_l) \qquad (20.38)
\end{aligned}
$$

其中，在式(20.38)中，我们取消了 (X_i,I_l) 的边缘概率密度函数，从而使它变成了一个无条件累积分布函数对一个概率密度函数的比率。

假定 X_l 有 q 个连续成分和 r 个离散成分，我们可以通过式(20.34)估计 $G(b,x,i)$ 和 $g(b,x,i)$。

648 当拍卖是异质时，Guerre 等(2000)建议通过下式估计 $G(\cdot)$ 和 $g(\cdot)$：

$$
\begin{aligned}
\tilde{G}(b,x,i) &= \frac{1}{L}\sum_{l=1}^{L}\frac{1}{I_l}\sum_{i=1}^{I}\mathbf{1}(B_{il}\leqslant b)K_h(X_l,x)L(I_l,i,\lambda) \\
\tilde{g}(b) &= \frac{1}{L}\sum_{l=1}^{L}\frac{1}{I_l}\sum_{i=1}^{I_l}k_{h_b}(B_{il},b)K_h(X_l,x)L(I_l,i,\lambda) \qquad (20.39)
\end{aligned}
$$

私人价值 V_{il} 通过下式来估计：

$$
\hat{V}_{il} = B_{il} - \frac{1}{I_l-1}\hat{\psi}(B_{il},X_l,I_l) \qquad (20.40)
$$

其中 $\hat{\psi}(b,x,i)=\tilde{G}(n,x,i)/\tilde{g}(b)$。然而，$\tilde{\psi}$ 在 (B,X,I) 支撑的边界是 ψ 的一个渐近有偏估计量。Guerre 等进一步介绍了一个在支撑边界附近可以调整观测

值的修剪函数;关于调整函数的详细情况参见 Guerre 等。

20.3.2　条件独立的私人信息拍卖

Li, Perrigen 和 Vuong(2000)将 IPV 模型扩展到条件独立私人信息(conditionally independent private information ,CIPI)的情况。令 σ_i 表示投标者 $i(i=1,\cdots,n)$的私人价值,v 表示共同价值。令 $F_\sigma(\cdot)$和 $F_{\sigma|v}(\cdot|\cdot)$分别表示 σ 的累积分布函数和给定 v 下 σ 的条件累积分布函数,有相对应的概率密度函数 $f_\sigma(\cdot)$和$f_{\sigma|v}(\cdot|\cdot)$以及非负支撑$[\underline{v},\bar{v}]$和$[\underline{\sigma},\bar{\sigma}]$。CIPI 模型假定投标者的信号 σ_i 在给定共同价值 v 下是条件独立的,因此

$$f(\sigma_1,\cdots,\sigma_n,v)=f_v(v)\prod_{i=1}^{n}f_{\sigma|v}(\sigma_i\mid v) \tag{20.41}$$

Li 等(2000)考虑了严格递增可微对称贝叶斯纳什均衡策略的情况。参与者 i 选择他的出价 b_i 来最大化

$$E[(U_i-b_i)\mathbf{1}(B_i\leqslant b_i)\mid\sigma_i]$$

其中

$$B_i=s(Y_i),\quad Y_i=\max_{j\neq i}\sigma_j$$

并且 $s(\cdot)$是均衡策略。一阶条件是

$$s^{(1)}(\sigma_i)=\frac{[V(\sigma_i,\sigma_i)-s(\sigma_i)]f_{Y|\sigma}(\sigma_i\mid\sigma_i)}{F_{Y|\sigma}(\sigma_i\mid\sigma_i)} \tag{20.42}$$

其中　　649

$$V(\sigma_i,Y_i)=E[U(\sigma_i,v)\mid\sigma_i,Y_i]$$

$F_{Y|\sigma}(\cdot|\cdot)$表示给定 σ_i 下 Y_i 的条件累积分布函数,$f_{Y|\sigma}(\cdot|\cdot)$是它的概率密度函数。

当保留价格是非粘性时,解被给出为

$$b_i=s(\sigma_i)=V(\sigma_i,\sigma_i)-\int_{\underline{\sigma}}^{\sigma_i}L(\alpha\mid\sigma_i)\mathrm{d}V(\alpha,\alpha) \tag{20.43}$$

其中

$$L(\alpha\mid\sigma_i)=\exp\left[-\int_{\underline{\sigma}}^{\sigma_i}f_{Y|\sigma}(u\mid u)/F_{Y|\sigma}(u\mid u)\mathrm{d}u\right]$$

在这个模型中,假定 $U(\sigma_i,v)=\sigma_i$,从而每个投标者的私人信息是其个人的效用函数。为了可识别,Li 等(2000)假定 $\sigma_i=v\eta_i$,η_i 是独立同分布且均值等于 1,并且 v 和 η_i 是相互独立的。

令 $G_{B|b}(\cdot|\cdot)$和 $g_{B|b}(\cdot|\cdot)$分别表示给定 b_i 下 B_i 的条件累积分布函数和概率密度函数,则

$$G_{B|b}(X_i \mid X_i) = \mathrm{P}[B_i \leqslant X_i \mid b_i = X_i]$$
$$= \mathrm{P}[Y_i \leqslant s^{-1}(X_i) \mid \sigma_i = s^{-1}(X_i)]$$
$$= F_{Y|\sigma}(s^{-1}(X_i) \mid s^{-1}(X_i)) \tag{20.44}$$

且

$$g_{B|b}(X_i \mid X_i) = \frac{f_{Y|\sigma}(s^{-1}(X_i) \mid s^{-1}(X_i))}{s^{(1)}(s^{-1}(X_i))} \tag{20.45}$$

利用式(20.44)和式(20.45)以及 $\sigma = s^{-1}(b)$,则式(20.42)可以被写为

$$V(\sigma,\sigma) = b + \frac{G_{B|b}(b \mid b)}{g_{B|b}(b \mid b)} \equiv \xi(b,G) \tag{20.46}$$

可以使用在20.3.1节讨论过的那些类似的方法由下式来估计 σ_i:

$$\ln(\sigma_i) = \ln(v) + ln(\eta_i) \equiv \ln c + \ln \varepsilon_i$$

其中

$$\ln c = \ln v + E[\ln \eta_i], \quad \ln \varepsilon_i = \ln \eta_i - \mathrm{E}[\ln \eta_i]$$

基于可观测出价 $b_{il}, i = 1,\cdots,n, l = 1,\cdots,L$,我们通过 $\hat{G}_{B_1|b_1}(\cdot \mid \cdot)/\hat{g}_{B_1|b_1}(\cdot \mid \cdot)$来估计 $G_{B_1|b_1}(\cdot \mid \cdot)/g_{B_1|b_1}(\cdot \mid \cdot)$,其中

$$\hat{G}_{B_1,b_1}(B,b) = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}(B_{il} \leqslant B)K_{h_1}(b_{il},b)$$
$$\hat{g}_{B_1,b_1}(B,b) = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{n}\sum_{i=1}^{n}k_{h_2}(B_{il},B)k_{h_1}(b_{il},b) \tag{20.47}$$

650

利用式(20.46)我们通过下式估计 V_{il}:

$$\hat{V}_{il} = b_{il} + \frac{\hat{G}_{B|b}(b_{il} \mid b_{il})}{g_{B|b}(b_{il} \mid b_{il})} \equiv \hat{\xi}(b_{il})$$

如20.3.1节讨论的,为了避免边界效应,一些修剪是必需的。Li 等(2000)建议修剪掉那些对应在标价支撑$[0,b_{\max}]$边界 h_1 距离内标价的 $\hat{V}_{il}$值。

接下来,可以通过下面的式子在 n 个投标者中估计任何两个投标者的 $\ln V_{il}$ 的联合特征函数:

$$\hat{\psi}(u_1,u_2) = \frac{1}{n(n-1)}\sum_{1 \leqslant i \neq j \leqslant n}\frac{1}{L_T}\sum_{l=1}^{L_T}\exp(iu_1 \ln \hat{V}_{il} + iu_2 \ln \hat{V}_{jl})$$

接着可以通过下面的式子估计对 $x \in [\ln \underline{c}, \ln \bar{c}]$, $y \in [\ln \underline{\varepsilon}, \ln \bar{\varepsilon}]$,$c$ 和 ε 的概率密度函数:

$$\hat{f}_c(x) = \frac{1}{2\pi}\int_{-T}^{T}\mathrm{e}^{-itx}\hat{\phi}_c(t)\,\mathrm{d}t \tag{20.48}$$

$$\hat{f}_{\varepsilon}(y) = \frac{1}{2\pi}\int_{-T}^{T} e^{-ity}\hat{\phi}_{\varepsilon}(t)\,dt \tag{20.49}$$

其中 T 是一个平滑参数,并且

$$\hat{\phi}_{c}(t) = \exp\int_{0}^{t} \frac{\partial\hat{\psi}(0,u_2)/\partial u_1}{\hat{\psi}(0,u_2)}du_2 \tag{20.50}$$

$$\hat{\phi}_{\varepsilon}(t) = \hat{\psi}(t,0)/\hat{\phi}_{c}(t) \tag{20.51}$$

最后,我们通过下式来估计 $f_v(\cdot)$ 和 $f_\eta(\cdot)$:

$$\hat{f}_{v}(x) = \hat{f}_{c}(x + \hat{E}(\ln\eta)), \quad \hat{f}_{\eta}(x) = \hat{f}_{\varepsilon}(x - \hat{E}(\ln\eta)) \tag{20.52}$$

其中 $\hat{E}(\ln\eta) = -\ln\hat{E}(\varepsilon)$。

Li 等(2000)建立了 $\hat{f}_v(\cdot)$ 对 $f_v(\cdot)$ 以及 $\hat{f}_\eta(\cdot)$ 对 $f_\eta(\cdot)$ 在一些正则性条件下的均匀一致性,包括当 $L\to\infty$ 时,T 相应发散于 ∞。

一个重要的实际问题是怎样在投标者信息的私人和共同价值模型间进行选择。Haile, Hong 和 Shum(2003)发展了在一级价格封闭式投标拍卖中对共同价值的检验。他们的检验是非参数的,并且仅仅要求观测值是在每个拍卖被提交时的报价。在 Haile 等(2003)中报告的美国森林服务木材拍卖的实证应用表明混合证据不支持私人价值的假定。

20.4 多元分布的基于 copula 的半参数估计

651

最近,所谓的"copula"已经引起了许多当之无愧的关注,尤其来自那些工作在银行和保险行业的人。对于 copula 应用在金融方面的一个综述,参见 Embrechts, McNeil 和 Straumann(1999)。[①] 在本节我们讨论基于 copula 的半参数多元分布的估计。

20.4.1 copula 函数的一些背景

我们知道自 Sklar(1959)的开创性工作以来,具有连续边缘累积分布函数 $F_j(x_j)=P(X_j\leqslant x_j)$ $(j=1,\cdots,p)$ 的一个随机向量 $(X_1,\cdots,X_p)$ 的联合行为可以被与它相联的"copula 函数"C 唯一特征化,对于所有 $(v_1,\cdots,v_p)\in[0,1]^p$,该函数被定义为

$$C(v_1,\cdots,v_p) = P(F_1(X_1)\leqslant v_1,\cdots,F_p(X_p)\leqslant v_p) \tag{20.53}$$

---

① copula 被证明在微观计量中也是有用的。参见 Lee(1983)以及 Trivedi 和 Zimmer(2007)对 copula 在样本选择模型中的使用,Heckman 和 Honoré(1989)在竞争风险模型中的使用。

方程(20.53)表明 copula 函数本身是一个累积分布函数。此外,通过附录 A 的定理 A.2 可知,每一个成分 $V_j = F_j(X_j)$ 有一个均匀边缘分布。

令 $F(x_1,\cdots,x_p)$ 表示随机向量 $(X_1,\cdots,X_p)$ 的联合累积分布函数,则 copula 函数将边缘分布与联合累积分布函数连接如下:

$$F(X_1,\cdots,X_p) = C(F_1(X_1),\cdots,F_p(X_p)) \tag{20.54}$$

令 $f_j(\cdot)(j=1,\cdots,p)$ 和 $c(v_1,\cdots,v_p)$ 分别表示与 $F_j(\cdot)(j=1,\cdots,p)$ 和 $C(v_1,\cdots,v_p)$ 相对应的概率密度函数。令 $\mathcal{X}_j$ 表示 X_j 的支撑,对任何 $(x_1,\cdots,x_p) \in \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$,与式(20.54)给出的 $F(\cdot)$ 相对应的概率密度函数 $f(\cdot)$ 有如下的形式:

$$f(x_1,\cdots,x_p) = c(F_1(x_1),\cdots,F_p(x_p))\prod_{j=1}^{p} f_j(x_j) \tag{20.55}$$

或者等价地,

$$c(v_1,\cdots,v_p) = f(x_1,\cdots,x_p)/\prod_{j=1}^{p} f_j(x_j)$$

其中 $v_j = F_j(x_j)$, $j=1,\cdots,p$。

方程(20.55)表明 copula 函数完整刻画了随机向量 $(X_1,\cdots,X_p)$ 的依赖结构。很显然,当且仅当 $X_1,\cdots,X_p$ 相互独立时,$c(v_1,\cdots,v_p) \equiv 1$。

一个广泛使用的 copula 函数是高斯 copula。令 Φ 表示一元标准正态累积分布函数,令 $\Phi_{\Sigma,p}$ 表示一个具有零均值及相关矩阵 Σ 的 p 维正态累积分布函数,则具有相关矩阵 Σ 的 p 维高斯 copula 由下式给出:

$$C(\boldsymbol{v};\Sigma) = \Phi_{\Sigma,p}(\Phi^{-1}(v_1)\cdots,\Phi^{-1}(v_d))$$

其中 $\boldsymbol{v} = (v_1,\cdots,v_p)$。它的 copula 密度是

$$c(\boldsymbol{v};\Sigma) = \frac{1}{\sqrt{\det(\Sigma)}} \times \exp\left\{\frac{-(\Phi^{-1}(v_1)\cdots,\Phi^{-1}(v_p))'(\Sigma^{-1}-I_p)(\Phi^{-1}(v_1),\cdots,\Phi^{-1}(v_p))}{2}\right\}$$

类似地,令 T_ν 为自由度 $\nu>2$ 的标量标准 t 分布,令 $T_{\Sigma,\nu}$ 为具有相关矩阵 Σ 的 p 维标准 t 分布,则具有相关矩阵 Σ 的 p 维(标准)t-copula 是

$$C(\boldsymbol{v};\Sigma,\nu) = T_{\Sigma,\nu}(T_\nu^{-1}(v_1),\cdots,T_\nu^{-1}(v_d))$$

并且 t-copula 密度是

$$c(\boldsymbol{v};\Sigma,\nu) = \frac{\Gamma\left(\frac{\nu+p}{2}\right)\left[\Gamma\left(\frac{\nu}{2}\right)\right]^{p-1}}{\sqrt{\det(\Sigma)}\left[\Gamma\left(\frac{\nu+1}{2}\right)\right]^{p}}\left(1+\frac{x'\Sigma x}{\nu-2}\right)^{-\frac{\nu+p}{2}}\prod_{j=1}^{p}\left(1+\frac{x_i^2}{\nu-2}\right)^{\frac{\nu+1}{2}}$$

其中 $x=(x_1,\cdots,x_p)'$, $x_i = T_\nu^{-1}(v_i)$,关于各种各样的非高斯、非对称 copula 函数

可以参见 Joe(1997)和 Nelson(2006)。

20.4.2　半参数基于 copula 的多元分布

在本节我们讨论 Genest, Ghoudi 和 Pivest(1995)以及 Chen, Fan 和 Tsyrennikov(2006)等人所研究的半参数 copula 模型。这类分布由一个以非参数边缘分布评估的参数 copula 函数刻画。它有两点吸引人的性质:(i) 它允许人们为一个多元随机变量的依赖结构和边缘行为分开建模;(ii) 它避免了维数诅咒问题,因为它仅仅包含一维(未知)一元累积分布函数。

若选择一个参数 copula 函数,比如 $c(v_1,\cdots,v_p;\theta)$,其中 $\theta\in\Theta$ 是一个有限维参数(比如,Θ 是 $\mathbb{R}^q$ 的一个有界子集),并且利用它来取代式(20.55)中的 $c(v_1,\cdots,v_p)$(允许边缘分布 $f_j(\cdot)$ 未设定),则得到的多元概率密度函数 $f(x_1,\cdots,x_p)$ 仅仅包含一维非参数函数 $f_j(x_j)$, $j=1,\cdots,p$。于是达到了在非参数成分中维数减少的目的。存在很多可行的参数 copula 函数,且通过选择不同的参数 copula 函数,得到的联合概率密度函数 $f(\cdot)$ 可以是对称的或者非对称的,拥有各种尾部依赖性质等。接下来,我们转向式(20.55)给出的半参数模型的估计。

653

20.4.3　一个两步估计过程

一个简便的估计方法是由 Oakes(1994)和 Genest 等(1995)提出的两步估计过程,可以通过下式来估计 θ:

$$\tilde{\theta} = \arg\max_{\theta\in\Theta}\left[\sum_{i=1}^{n}\ln c(\tilde{F}_{n1}(X_{1i}),\cdots,\tilde{F}_{np}(X_{pi});\theta)\right] \tag{20.56}$$

其中下式:是 $F_j(x_j)$, $j=1,\cdots,p$ 的重设比例(rescaled)经验累积分布函数估计量

$$\tilde{F}_{nj}(x_j) = \frac{1}{n+1}\sum_{i=1}^{n}\mathbf{1}(X_{ji}\leqslant x_j)$$

注意这里的因子是 $1/(n+1)$ 而不是 $1/n$。这种重新调整避免了当一些 v_i 倾向于1时, $\ln\{c(v_1,\cdots,v_p;\theta)\}$ 的潜在无界引起的困难。Genest 等建立了以下总结的 $\tilde{\theta}$ 的 $\sqrt{n}$ 正态性。

对 $j=1,\cdots,p$,令

$$l(v_1,\cdots,v_p;\theta) = \ln\{c(v_1,\cdots,v_p;\theta)\},$$

$$l_{\theta,\theta} = \frac{\partial^2 \ln c(v_1,\cdots,v_p;\theta)}{\partial\theta\partial\theta'}$$

$$l_{\theta,j} = \frac{\partial^2 \ln c(v_1,\cdots,v_p,\theta)}{\partial v_j \partial \theta}$$

$$W_j(V_j) = \int_0^1 \cdots \int_0^1 [\mathbf{1}(V_j \leqslant v_1) - v_1]$$
$$\times l_{\theta,j}(v_1,\cdots,v_p,\theta) c(v_1,\cdots,v_p,\theta) \mathrm{d}v_1 \cdots \mathrm{d}v_p$$

Genest 等(1995)证明了如下的结论。

654

定理 20.2 在 Genest 等(1995)给出的正则性条件下,

$$\sqrt{n}(\tilde{\theta} - \theta) \to N(0,V) \text{ 依分布收敛}$$

其中 $V = B^{-1}\Sigma B^{-1}$, $B = -E[l_{\theta,\theta}(F_1(X_1),\cdots,F_p(X_p),\theta)]$,并且

$$\Sigma = \text{var}[l_\theta(F_1(X_1),\cdots,F_p(X_p),\theta) + \sum_{j=1}^{p} W_j(X_j)]$$

Genest 等(1995)也提出了对于 B 和 Σ 的一致估计量。B 通过下式来估计:

$$\tilde{B} = -\frac{1}{n}\sum_{i=1}^{n} l_{\theta,\theta}(\tilde{F}_{n1}(X_{1i}),\cdots,\tilde{F}_{np}(X_{ni});\tilde{\theta})$$

重设比例经验 copula 函数可以通过下式来估计:

$$\tilde{C}_n(v_1,\cdots,v_p) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{p}\mathbf{1}(\tilde{F}_{nj}(X_{ji}) \leqslant v_j)$$

令 $\tilde{U}_i = l_\theta(\tilde{F}_{n1}(X_{1i}),\cdots,\tilde{F}_{np}(X_{ni}),\tilde{\theta})$,并且令

$$\tilde{V}_t = \tilde{U}_t - \sum_{j=1}^{p}\int \mathbf{1}(\tilde{F}_{nj}(X_{ji}) \leqslant v_j)$$
$$\times l_\theta(v_1,\cdots,v_p,\tilde{\theta}) l_j(v_1,\cdots,v_p,\tilde{\theta}) \mathrm{d}\tilde{C}_n(v_1,\cdots,v_p)$$
$$= \tilde{U}_t - n^{-1}\sum_{i=1}^{n}\sum_{j=1}^{p}\int \mathbf{1}[\tilde{F}_{nj}(X_{ji}) \leqslant \tilde{F}_{nh}(X_{ji})]$$
$$\times l_\theta(\tilde{F}_{h1}(X_{j1}),\cdots,\tilde{F}_{np}(X_{np}),\tilde{\theta})$$
$$\times l_j(\tilde{F}_{n1}(X_{j1}),\cdots,\tilde{F}_{np}(X_{jp}),\tilde{\theta})$$

则 Σ 可以通过 $\{\tilde{V}_t\}_{t=1}^{n}$ 的样本方差来估计,即

$$\hat{\Sigma} = n^{-1}\sum_{i=1}^{n} A_{ni}A_{ni}'$$

其中

$$A_{ni} = \tilde{V}_i - n^{-1}\sum_{t=1}^{n}\hat{V}_t$$

以上讨论的两步估计过程的计算是简单的,然而,一般都不是有效的。我

们现在讨论一个有效估计过程。

20.4.4　一步有效估计过程

Chen 等(2006)提出了一个一步基于筛的有效估计方法。他们建议使用一个线性筛(序列)来近似边缘概率密度函数的平方根,即 655

$$\mathcal{F}_{nj} = \left\{ f_{K_{nj}}(x) = \left[\sum_{k=1}^{K_{nj}} a_k A_k(x) \right]^2, \int f_{K_{nj}}(x)\,\mathrm{d}x = 1 \right\} \tag{20.57}$$

其中 $K_{nj} \to \infty$ 且 $K_{nj}/n \to 0$,其中 $\{A_k(\cdot): k \geqslant 1\}$ 包含了已知的基础函数,且 $\{a_k: k \geqslant 1\}$ 是未知筛系数的集合。

若 $\mathcal{X}_j = [0,1]$,则近似基础函数 $A_k(\cdot)$ 可以成为 B 样条筛 $\mathrm{Spl}(K_n)$,或者多项式筛

$$\mathrm{Pol}(K_n) = \left\{ \sum_{k=1}^{K_n} a_k x^k, x \in [0,1]: a_k \in \mathbb{R} \right\}$$

或者三角筛

$$\mathrm{TriPol}(K_n) = \left\{ a_0 + \sum_{k=1}^{K_n} [a_k \cos(k\pi x) + b_k \sin(k\pi x)], x \in [0,1]: a_k \in \mathbb{R} \right\}$$

若 $\mathcal{X}_j = \mathbb{R}$,那么如果概率密度函数 $f_j(\cdot)$ 接近于指数薄尾,可以使用在 Gallant 和 Nychka (1987)中给出的 Hermite 多项式筛:

$$\mathcal{F}_{nj} = \left\{ f_{K_{nj}(x)} = \sigma^{-1} \left[\varepsilon_0 + \left\{ \sum_{k=1}^{K_{nj}} a_k \left(\frac{x - \zeta_0}{\sigma} \right)^k \right\}^2 \right] \right.$$
$$\left. \times \exp\left\{ -\frac{(x-\zeta_0)^2}{2\sigma^2} \right\}: \int f_{K_{nj}}(x)\,\mathrm{d}x = 1 \right\} \tag{20.58}$$

其中 $\varepsilon_0 > 0, \sigma > 0, \zeta_0$ 和 $a_k \in \mathbb{R}$。

若另一方面,$f_j(\cdot)$ 有多项式厚尾,可以使用样条小波筛:

$$\mathcal{F}_{nj} = \left\{ f_{K_{nj}}(x) = \left[\sum_{k=0}^{K_{nj}} \sum_{l \in K_n} a_{kl} 2^{k/2} B_\gamma(2^k x - l) \right]^2, \int f_{K_{nj}}(x)\,\mathrm{d}x = 1 \right\} \tag{20.59}$$

其中 $B_\gamma(\cdot)$ 表示 γ 阶的基本 B 样条,在第 15 章式(15.2)中给出。

令 $\alpha = (\theta', f_1, \cdots, f_p)'$ 并且令

$$l(\alpha, Z_i) = \ln\left\{ c(F_1(X_{1i}), \cdots, F_p(X_{pi}); \theta) \prod_{j=1}^{p} f_j(X_{ji}) \right\}$$

同样,令 656

$$\mathcal{A}_n = \Theta \times \prod_{j=1}^{p} \mathcal{F}_{nj}$$

其中线性筛空间可以是如上讨论过的筛之一。Chen 等(2006)提出了 α 的如下筛估计量：

$$\hat{\alpha} = \arg\max_{\alpha \in \mathcal{A}_n} \sum_{i=1}^{n} l(\alpha, Z_i) \tag{20.60}$$

$F_j(\cdot)$的最终估计量由下式获得：

$$\hat{F}_{nj}(x_j) = \int \mathbf{1}(y \leqslant x_j) \hat{f}_{hj}(y) \mathrm{d}y$$

Chen 等(2006)证明 $\hat{\alpha}$ 是半参数有效的。更具体地，他们还证明

$$\sqrt{n}(\hat{\theta} - \theta) \to N(0, I_0(\theta)^{-1}) \text{ 依分布收敛}$$

其中 $I_0(\theta)^{-1}/n$ 是任何正则估计量的渐近方差的半参数下界。但是，$I_0(\theta)$一般没有一个封闭形式的表达式。然而，Chen 等能够为 $I_0(\theta)$ 构建一个如下形式的一致估计量。令

$$\hat{U}_i = (\hat{U}_{1i}, \cdots, \hat{U}_{pi})' = (\hat{F}_{n1}(X_{1i}), \cdots, \hat{F}_{np}(X_{pi}))'$$

$I_0(\theta)$的一个一致估计量被给定为

$$\hat{I}_0(\theta) = \min_{g_j \in \mathcal{A}_n, j=1,\cdots,p} n^{-1} \sum_{i=1}^{n} \{ B_{ni}(\hat{\theta}, \hat{U}_1, \cdots, \hat{U}_p)' B_{ni}(\hat{\theta}, \hat{U}_1, \cdots, \hat{U}_p) \} \tag{20.61}$$

其中

$$B_{ni}(\hat{\theta}, \hat{U}_1, \cdots, \hat{U}_p) = \frac{\partial \ln c(\hat{U}_i, \hat{\theta})}{\partial \theta} - \sum_{j=1}^{p} \left[\frac{\partial \ln c(\hat{U}_i, \hat{\theta})}{\partial v_j} \int_0^{\hat{U}_{ji}} g_j(v) \mathrm{d}v + g_j(\hat{U}_{ji}) \right]$$

Chen 等证明 $\hat{I}_0(\theta) - I_0(\theta) = o_p(1)$。注意式(20.61)要求一个最优过程，因为 $I_0(\theta)$没有一个封闭形式的表达式。

边缘累积分布函数可以通过下式来估计：

$$\hat{F}_{nj}(x_j) = \int_{-\infty}^{x_j} \hat{f}_{nj}(v) \mathrm{d}v$$

Chen 等进一步证明得到的估计量 $\hat{F}_{nj}(x_j)$是$\sqrt{n}$一致的，并且对于 $F_j(x_j)$是半参数有效估计量，即

$$\sqrt{n}(\hat{F}_{nj}(x_j) - F_j(x_j)) \to N(0, V_{0j}(x_j)) \text{ 依分布收敛}$$

然而，一般 $V_{0j}(x_j)$没有一个封闭形式的表达式。然而，仍然可以获得一个

基于筛的一致估计量，比如 $\hat{V}_{0j}(x_j)$ 使得

$$\hat{V}_{0j}(x_j) = V_{0j}(x_j) + o_p(1)$$

参见 Chen 等对于 $\hat{V}_{0j}(x_j)$ 的一个具体的定义。当 $X_1,\cdots,X_p$ 相互独立时（在这种情况下 $c(v_1,\cdots,v_p;\theta)\equiv 1$），则

$$V_{j0}(x_j) = F_j(x_j)(1 - F_j(x_j))$$

有一个封闭形式的表达式，且与 $\sqrt{n}\tilde{F}_{nj}(x_j)$（$F_j(x_j)$ 的经验估计量）的渐近方差相吻合。对于更普遍的情况，我们有

$$V_{0j}(x_j) \leqslant F_j(x_j)(1 - F_j(x_j))$$

20.4.3 节和 20.4.4 节的估计过程都假定所选择的参数 copula 函数是正确设定的。在实践中，需要检验一个所选择的参数 copula 是否确实足够接近于未知的真实 copula 函数。在接下来的一节我们讨论检验一个参数 copula 函数的问题。

20.4.5　检验一个 copula 的参数函数形式

令 $\{C_0(v_1,\cdots,v_p;\theta):\theta\in\Theta\}$ 是关于 $\boldsymbol{X}=(X_1,\cdots,X_p)$ 的一类参数 copula。我们有兴趣检验如下的原假定：

$$H_0:P[C(v_1,\cdots,v_p) = C_0(v_1,\cdots,v_p;\theta_0)] = 1,\text{某些 } \theta_0 \in \Theta$$

其中 Θ 是在 $\mathbb{R}^q$ 上的一个有界子集。备择假定是 H_0 的否定，即

$$H_1:P[C(v_1,\cdots,v_p) = C_0(v_1,\cdots,v_p;\theta)] < 1$$

令 $C_{0j}(v_1,\cdots,v_p;\theta_0)$ 表示在 H_0 下 $V_1,\cdots,V_j$ 的联合累积分布函数，即

$$C_{0j}(v_1,\cdots,v_j;\theta_0) = C_0(v_1,\cdots,v_j,1,\cdots,1;\theta_0)$$

（对 $l>j,V_l=1$）。同样，令 $C_{0j}(V_j;\theta_0\mid V_1,\cdots,V_{j-1})$ 表示在 H_0 下给定 $(V_1,\cdots,V_{j-1})$ 时 V_j 的条件累积分布函数。则可知

$$C_{0j}(v_j;\theta_0 \mid v_1,\cdots,v_{j-1}) = \frac{\partial^{j-1}C_{0j}(v_1,\cdots,v_j;\theta_0)}{\partial v_1\cdots\partial v_{j-1}} \times \frac{\partial^{j-1}C_{0,j-1}(v_1,\cdots,v_{j-1};\theta_0)}{\partial v_1\cdots\partial v_{j-1}}$$

如下定义新的随机变量 Z_j。首先，令 $Z_1=V_1$，然后对 $j=2,\cdots,p$，令

$$Z_j = C_{0j}(V_j;\theta_0 \mid V_1,\cdots,V_{j-1})$$

因为 copula 函数是一个多元的累积分布函数，由 Rosenblatt(1952) 可知当且仅当 $Z_1,\cdots,Z_p$ 是独立同分布 $U[0,1]$ 的随机变量时，H_0 成立。因此，原假定可以被等价地写为 H_0：$P(g(Z_1,\cdots,Z_p)=1)=1$，其中 $g(z_1,\cdots,z_p)$ 是 $Z_1,\cdots,Z_p$ 的联合概率密度函数。可以通过下式来估计 $g(z_1,\cdots,z_p)$：

$$\tilde{g}(z_1,\cdots,z_p) = \frac{1}{nh_1\cdots h_p}\sum_{t=1}^{n}[k_h(\hat{Z}_t,z)]$$

658 其中

$$\bar{k}_h(\hat{Z}_t,z) = \prod_{j=1}^{p} k_{hj}(\hat{Z}_{jt},z_j)$$

且 $k_h(x,y)$ 是在 20.1.4 节式中(20.19)中定义的一元边界修正核函数(具有支撑$[-1,1]$,在 20.2.4 节的式(20.19)中被定义)。令 $\hat{Z}_{1t}=\hat{F}_1(\hat{Y}_{1t})$,并且对 $j=2,\cdots,p$,令

$$\hat{Z}_{jt} = C_{0j}(\hat{F}_j(Y_{jt});\hat{\theta}\mid \hat{F}_1(Y_{1t}),\cdots,\hat{F}_{j-1}(Y_{j-1,t}))$$

其中 $\hat{\theta}$ 是在原假定下 θ_0 的一个$\sqrt{n}$一致估计(不需要是有效估计量)。这个检验统计量将基于

$$\int_0^1\cdots\int_0^1[\hat{g}(z_1,\cdots,z_p)-1]^2\mathrm{d}z_1\cdots\mathrm{d}z_p \tag{20.62}$$

注意式(20.62)包含两个求和(因为 $\hat{g}(\cdot)$ 包含一个求和),即 $\sum_{t=1}^{n}\sum_{s=1}^{n}$,若我们移掉 $t=s$ 项,我们将获得如下的检验统计量 $\left(\int_0^1\mathrm{d}z=\int_0^1\cdots\int_0^1\mathrm{d}z_1\cdots\mathrm{d}z_p\right)$:

$$\begin{aligned}\hat{I}_n &= \frac{1}{n(n-1)}\sum_{t=1}^{n}\sum_{s\neq t}^{n}\int_0^1 k_h(\hat{Z}_t,z)k_h(\hat{Z}_t,z)\mathrm{d}z\\ &\quad -\frac{2}{n}\sum_{t=1}^{n}\int_0^1 k_h(\hat{Z}_t,z)\mathrm{d}z+1\end{aligned} \tag{20.63}$$

Chen 等(2006)已经证明了接下来的结论。

定理 20.3 在 Chen 等(2006)给出的正则性条件下,以及在 H_0 下,

$$\hat{T}_n = n(h_1\cdots h_p)^{1/2}\hat{I}_n/\sigma_{n,0}\rightarrow N(0,1) \text{ 依分布收敛}$$

其中

$$\sigma_{n,0}^2 = 2\left\{\int_{-1}^{1}\left[\int_{-1}^{1}[k(v)k(u+v)\mathrm{d}v]^2\mathrm{d}u\right\}^p$$

注意,不像 Chen 等(2006)考虑的原始检验统计量,定理 20.3 中的 $\hat{T}_n$ 检验没有非零的中心项,因为在式(20.63)中我们已经移除了 $i=j$ 项。

如上的 $\hat{T}_n$ 要求一个多元概率密度函数 $g(z_1\cdots,z_p)$ 的估计。当 p 较大时,将会遭遇维数诅咒问题。Chen 等(2006)也提出了另外一个仅包含一元概率密度函数估计的检验。它基于这样的事实,在 H_0 下,$W=\sum_{j=1}^{p}[\Phi^{-1}(Z_j)]^2$ 有一个χ_p^2分布。令 $F_{\chi_p^2}(w)$ 表示一个 χ_p^2 随机变量的累积分布函数,则 $F_{\chi_p^2}(W)$ 有 $U[0,1]$

分布。这激发了 Chen 等去构建一个基于$\int_0^1[\hat{g}_W(w)-1]^2\mathrm{d}w$ 的检验，其中

659

$$\hat{g}(w)=\frac{1}{nh}\sum_{t=1}^{n}k_h(w,F_{\chi_p^2}(\hat{W}_t))$$

$$\hat{W}_t=\sum_{j=1}^{p}[\Phi^{-1}(\hat{Z}_{jt})]^2$$

如同推导 $\hat{I}_n$ 检验时的情况，通过移除 $t=s$ 的项我们获得了如下的检验统计量：

$$\hat{I}_{n,2}=\frac{1}{n(n-1)}\sum_{t=1}^{n}\sum_{s\neq t}^{n}\int_0^1\int_0^1 k_h(F_{\chi_p^2}(\hat{W}_t),w)\kappa_h(F_{\chi_p^2}(\hat{W}_s),w)\mathrm{d}w$$

$$-\frac{2}{n}\sum_{t=1}^{n}\int k_h(F_{\chi_p^2}(\hat{W}_t),z)\mathrm{d}z+1$$

类似于定理 20.3 的结论，Chen 等(2006)表明，在 H_0 下，

$$nh^{1/2}\hat{I}_{n,2}/\sigma_{n,2}\to N(0,1)\text{ 依分布收敛}\tag{20.64}$$

其中

$$\sigma_{n,2}^2=2\int_{-1}^{1}\left[\int_{-1}^{1}k(u+v)k(v)\mathrm{d}v\right]^2\mathrm{d}u$$

copula 函数也被用来模型化时间序列回归模型的依赖结构。Chen 和 Fan (2006)讨论了半参数 copula 时间序列回归模型的估计。也可以参见 Brendstrup 和 Paarsch(2004)，他们考虑了拍卖背景的一个应用，并且使用了 Frank copula。

20.5　一个半参数转换模型

在本节我们考虑一个如下形式的转换模型：

$$\Lambda_0(Y)=X'\beta+u\tag{20.65}$$

其中 Y 是一个标量因变量，$\Lambda_0(\cdot)$是一个严格递增的函数，X 是解释变量向量，β 是系数向量，u 是独立于 X 的不可观测误差项，其具有一个累积分布函数记为 $F(\cdot)$。模型(20.65)根据 Λ_0 和 $F(\cdot)$的不同形式可以产生许多不同的模型，包括 Box-Cox 转换模型，加速衰变时间模型，比例风险模型以及混合比例风险模型。在参数背景下，需要设定 Λ_0 和 $F(\cdot)$，其中任何一个误设都会导致非一致估计。在本节我们讨论式(20.65)的非参数估计，其没有对 Λ_0 和 $F(\cdot)$施加函数形式限制。β 的一个$\sqrt{n}$一致估计可以通过把式(20.65)看成一个单指数模型而获得，这在第 8 章中有所涉及；参见 Han(1987)，Härdle 和 Stoker(1989)，Ichimura(1993)，Powell 等(1989)，以及 Sherman(1993)对于估计 β 的不同方法。在本节我们仅仅讨论 Λ_0 的估计。

660

大多数现有的非参数方法是两步非参数平滑方法。例如，Horowitz(1996)

以及 Ye 和 Duan(1997)提出基于条件累积分布函数的非参数估计来估计 Λ_0 和 $F(\cdot)$。Chen(2002)提出了一种避免非参数平滑技术的秩估计方法(即 Han(1987))。我们下面简单讨论 Horowitz 和 Chen 的方法。

需要一些识别条件来识别 Λ_0 和 β。我们假定对某些 y_0, $\Lambda_0(y_0)=0$,并且 $|\beta_1|=1$,其中 β_1 是 X 第一成分的系数,它条件于剩余变量的概率分布关于 Lebesgue 测度是绝对连续的。令 $\{X_i,Y_i\}_{i=1}^n$ 是(X,Y)的一个随机样本,$G(\cdot|z)$ 是 Y 条件于 z 的累积分布函数,其中 $z=x'\beta$。那么很容易证明

$$\begin{aligned} G(y\mid z) &= P[Y\leqslant y\mid z] \\ &= P[T(Y)\leqslant T(y)\mid z] \\ &= P[U+z\leqslant T(y)] \\ &= F[T(y)-z] \end{aligned}$$

Horowitz(1996)进一步证明

$$\Lambda(y) = -\int_{y_0}^{y}[G_y(v\mid z)/G_z(v\mid z)]\mathrm{d}v$$

其中

$$G_y(v\mid z) = \partial G(y\mid z)/\partial y,\quad G_z(v\mid z) = \partial G(y\mid z)/\partial z$$

因此,Horowitz(1996)建议通过下面的式子估计 $\Lambda(\cdot)$:

$$\Lambda_n(y) = -\int_{y_0}^{y}\int_{S_w} w(z)[G_{ny}(v\mid z)/G_{nz}(v\mid z)]\mathrm{d}z\mathrm{d}v \tag{20.66}$$

其中 w 是满足 $\int_{S_w} w(z)\mathrm{d}z = 1$ 的一个加权函数,$G_{ny}(v|z)$ 和 $G_{nz}(v|z)$ 分别是 $G_y(v|z)$ 和 $G_z(v|z)$ 的核估计。在一些正则性条件下,Horowitz 证明 $\sqrt{n}(\Lambda_n(\cdot)-\Lambda_0(\cdot))$ 收敛于一个零均值的高斯过程,其有一个定义良好的协方差函数。

661

接下来,我们转向 Chen(2002)的估计过程。定义 $d_{iy}=\mathbf{1}(Y_i\geqslant y)$ 并令 $d_{i0}=\mathbf{1}(Y_i\geqslant 0)$。根据 $\Lambda_0(\cdot)$ 的单调性质,我们有

$$d_{iy} = \mathbf{1}\{X_i\beta + u_i \geqslant \Lambda_0(y)\}$$

以及

$$E(d_{iy}\mid X_i) = 1 - F(\Lambda_0(y) - X_i\beta)$$

因此,对 $i\neq j$ (且 $\Lambda_0(y_0)=0$)

$$E[(d_{iy}-d_{iy_0})\mid X_i,X_j] \geqslant 0,当(X_i-X_j)\beta \geqslant \Lambda_0(y) \tag{20.67}$$

对于一个给定的 β 的初始估计,我们表示为 $\hat{\beta}$,Chen(2002)提出通过如下的式子估计 $\Lambda(y)$:

$$\hat{\Lambda}_n(y) = \arg\min_{\Lambda\in M_\Lambda}\frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n}(d_{iy}-d_{jy_0})\mathbf{1}(X_i\hat{\beta}-X_j\hat{\beta}\geqslant\Lambda) \tag{20.68}$$

其中 M_Λ 是一个适当的紧集。

Chen(2002)证明 $\hat{\Lambda}_n(y)$ 对于 $\Lambda_0(y)$ 是一个 $\sqrt{n}$ 一致估计量，并且 $\sqrt{n}(\hat{\Lambda}_n(\cdot)-\Lambda_0(\cdot))$ 弱收敛于一个零均值的高斯过程，且其有一个定义良好的协方差函数。

转换模型(20.65)已经被扩展到审查数据的情况，即 $\Lambda_0(y^*)=X'\beta+u$，其中人们仅仅观测到 $y=\min\{Y^*,C\}$ 以及 $\delta=\mathbf{1}(Y^*\leqslant C)$，而不是 Y^*。Gorgens 和 Horowitz(1999)提出了对于审查转换模型的一个基于核的方法，而 Chen(2002)同样将他的秩估计方法扩展到审查数据的情况。

Khan 和 Tamer(2007)考虑了在 Ridder(1990)中引入的如下广义加速衰退时间(GAFT)模型，

$$\Lambda(Y_i)=X_i'\beta_0+\varepsilon_i,\quad i=1,\cdots,n \tag{20.69}$$

其中 $(Y_i,X_i')'$ 是 $(q+1)\times 1$ 维的向量，Y_i 是因变量(比如，在接受治疗后的存活时间)，X_i 是可观测协变量的一个向量，$\Lambda(\cdot)$ 是单调但未知的函数。我们观察 (v_i,d_i) 而不是 Y_i，其中 v_i 是一个标量变量，d_i 是一个二元随机变量(即一个"右审查"转换模型)，即

$$\begin{aligned}\Lambda(v_i)&=\min(X_i'\beta_0+\varepsilon_i,c_i)\\ d_i&=\mathbf{1}(X_i'\beta_0+\varepsilon_i\leqslant c_i)\end{aligned} \tag{20.70}$$

其中，如果数据是非审查的，$v_i=Y_i$，对于审查观测值，$v_i=c_i$。Khan 和 Tamer 提出了一个类似于 Han(1987)的秩回归估计方法，并且他们建立了 β_0 最终估计量的渐近正态性。

20.6 习　题

662

习题 20.1　推导式(20.24)。

习题 20.2　定义

$$Z_{n2,t}=k_h(X_{t\Delta}-x)\int_{t\Delta}^{(t+1)\Delta}\mu(X_u)\,\mathrm{d}u$$

$$\hat{m}_4(x)=(n\Delta)^{-1}\sum_{t=1}^{n}Z_{n2,t}$$

然后我们有

$$\hat{m}_1(x)=\hat{m}_3(x)+\hat{m}_4(x)$$

其中我们使用了在定理 20.1(i)中证明的事实。

$$\operatorname{var}(\hat{m}_1(x))=\operatorname{var}(\hat{m}_3(x))+(\text{s.o.})=O((nh\Delta)^{-1})$$

证明这的确是真的，即证明 $\operatorname{var}(\hat{m}_4(x))$ 是 $O((nh)^{-1})$ 阶的，其小于 $\operatorname{var}(\hat{m}_3(x))$ 是 $O((nh\Delta)^{-1})$ 的阶 $\operatorname{var}(\hat{m}_3(x))$。

附录 A　背景统计概念

在文献中建立的很多理论结果使用了许多基本的

663

统计概念。下面我们给出在先前章节中推导结论所需要的一些核心概念。

1.1　概率、测度和可测空间

定义 A.1（随机实验）　一个随机实验是一个行为或观测，其结果在发生前是不确定的。一个流行的随机实验是“抛硬币”，因为我们在抛之前不能肯定预测硬币是正面朝上还是背面朝上。

定义 A.2（样本空间 Ω）　样本空间用 Ω 表示，其被定义为一个随机实验中所有可能结果的集合。

例 A.1　若一个硬币掷两次，那么样本空间是

$$\Omega = \{HH, HT, TH, TT\} = \{w_1, w_2, w_3, w_4\}$$

其中 $w_1 = HH$，$w_2 = HT$（即第一次抛的是 H，第二次是 T）。在这个例子中，有四个基本事件 $w_i, i = 1, \cdots, 4$。

664

定义 A.3（可数和不可数集）　令 $\mathbb{N}_+ = \{1, 2, 3, \cdots\}$ 表示自然数的集合。若 B 中的所有元素可以被安排成与 N_+ 中的元素一一对应，则称集合 B 是可数的。若一个集合有一个可数的子集，但它本身是不可数的，我们说这个集合是不可数的。

例 A.2　$\mathcal{B} = \{1, 3, 5, \cdots\}$ 是可数的，因为 $B = \{b_n\}_{n=1}^{\infty}$，其中 $b_n = 2n - 1$，那么有一个 1—1 映射：$n \to 2n - 1$ 从 $\mathbb{N}_+$ 到 B。类似地，有理数集也是可数的。但是，集合 $[0,1] \in \mathbb{R}$ 是不可数的，因为从 $\mathbb{N}_+$ 到 $[0,1]$ 找不到一个 1—1 映射。但是，在 B 中存在的子集是可数的。

例如，$A=\{1/n\}_{n=1}^{\infty}$是[0,1]的一个子集，它是可数的。

所有幂函数的集合$\{x^n\}_{n=0}^{\infty}=\{1,x,x^2,\cdots\}$有可数的（无限）函数，然而所有定义域为[0,1]的有界连续函数的集合是不可数的。

定义 A.4（被定义在 Ω 中的 σ 域） 令$\mathcal{F}$（其至少包含一个非空的集合）是 Ω 子集的一个集合，如果如下的两个条件成立，那么$\mathcal{F}$被称为一个 σ 域（或者 σ 代数）：

(i) 若 $A\in\mathcal{F}$，那么 $A^c\in\mathcal{F}$，其中 A^c 是 A 的补集。

(ii) 若 $A_1,A_2,\cdots,\in\mathcal{F}$，那么 $\bigcup_{i=1}^{\infty}A_i\in\mathcal{F}$。

如上的(i)和(ii)告诉我们一个 σ 域在补集和可数并运算下是闭合的。

很容易证明(a) $\Omega\in\mathcal{F}$，(b) $\varnothing\notin\mathcal{F}$（$\varnothing$表示空集），(c) 若 $A_1,A_2,\cdots,\in\mathcal{F}$，那么 $\bigcap_{i=1}^{\infty}A_i\in\mathcal{F}$。

(a)的证明：令 A 是$\mathcal{F}$的非空集合。那么根据上面的(i)有 $A^c\in\mathcal{F}$，因此根据(ii) 有 $\Omega=A\cup A^c\in\mathcal{F}$。 □

(b)的证明：根据(i)以及 $\Omega\in\mathcal{F}$的事实可得出$\varnothing=\Omega^c\in\mathcal{F}$。 □

(c)的证明：我们知道根据(i)和(ii)的性质有 $\left(\bigcup_{i=1}^{\infty}A_i^c\right)^c\in\mathcal{F}$。因此，根据 DeMorgan 定律， □

$$\bigcap_{i=1}^{\infty}A_i=\left(\bigcup_{i=1}^{\infty}A_i^c\right)^c\in\mathcal{F}$$

$(\Omega,\mathcal{F})$被称为一个可测空间。

例 A.3

(i) 若$\{\mathcal{F}\}=\{\varnothing,\Omega\}$是一个仅包含空集和整个样本空间的 σ 域，那么它被称为“微小的 σ 域”（一个显然令人不感兴趣的东西）。

(ii) 若 Ω 包含 n（其中 n 是有限的）个基本事件（“元素”），那么它总共有 2^n 个子集。包含所有 Ω 子集的集合（称为一个“幂集”）是一个 σ 域，并且是在 Ω 中被定义的最大的 σ 域。 665

定义 A.5（DeMorgan 定律） 令 $A_1,A_2,\cdots,\in\mathcal{F}$（$\mathcal{F}$是在 Ω 中被定义的 σ 域）。令 A_i^c 表示 A_i 的补集，那么

$$\left(\bigcup_{i=1}^{\infty}A_i\right)^c=\bigcap_{i=1}^{\infty}A_i^c$$

例 A.4 可以使用一个 Venn 图（很容易）直观地验证 DeMorgan 定律。例如，令 A_1表示一个中学篮球队员的集合，A_2是足球队员的集合。那么 $\bigcup_{i=1}^{2}A_i=$

$A_1 \cup A_2$是足球队或者篮球队学生的集合,因此,它的补集 $\left(\bigcup_{i=1}^{2} A_i\right)^c$ 是那些既不是足球队又不是篮球队队员的学生。DeMorgan 定律断言这个集合和 $\bigcap_{i=1}^{2} A_i^c = A_1^c \cap A_2^c$是相同的,它是那些不是篮球队成员($A_1^c$)以及不是足球队成员的学生的集合,这些当然是正确的。

定义 A.6($\sigma(\mathcal{A})$,$\mathcal{A}$生成的 σ 域) 令$\mathcal{A}$是 Ω 子集的一个非空集合。那么$\mathcal{A}$生成的 σ 域表示为 $\sigma(\mathcal{A})$,其满足

(i) $\mathcal{A} \in \sigma(\mathcal{A})$。

(ii) $\sigma(\mathcal{A})$是包含$\mathcal{A}$的最小的 σ 域。

例 A.5 令 A 是 Ω 的一个非空真子集,那么$\{A, A^c, \varnothing, \Omega\}$是一个 σ 域,且是包含 A 的最小的 σ 域。

对于抛硬币的例子,若我们选择 $A = w_1 = HH$,那么$\{w_1, \{w_2, w_3, w_4\}, \varnothing, \Omega\}$是由 $A = w_1$ 生成的 σ 域。这个 σ 域没有包括$\{w_2\}$,$\{w_3\}$等。幂集(其包括所有 Ω 的子集)也包括 $A = w_1$,但它不是包含 A 的最小的 σ 域,所以幂集不是 $\sigma(A)$ $(A = w_1)$。

当样本空间包含$\mathbb{R}$ 的一个区间,比如说 $\Omega = \mathbb{R}$(它包含不可数的点),那么$\mathbb{R}$ 的所有子集的集合就太大了,并且在这种情况不能对$\mathbb{R}$ 的所有子集定义一个适当的测度。因此,对$\mathbb{R}$ 的子集施加一些约束是必要的。这可以通过要求$\mathbb{R}$ 的子集属于 Borel σ 域来做到,其被定义如下。

666

定义 A.7(Borel 域和 Borel 集) 令$\mathbb{R}$ 表示实数的集合,即$\mathbb{R} = \{y: -\infty < y < \infty\}$。令 $A_x = \{y: -\infty < y < x\} = (-\infty, x)$,且定义$\mathcal{A} = \{A_x, x \in \mathbb{R}\}$。那么$\mathcal{A}$生成的 σ 域被称为在$\mathbb{R}$ 上的 Borel(σ)域。

可以证明$\{a\}$,$[a,b]$,(a,b),$[a,b)$,$(a,b]$,$[a,+\infty)$都属于 Borel 域。因此,一个 Borel 域包括所有实际感兴趣的集合。

例 A.6 从$(-\infty, a) \in \sigma(\mathcal{A})$,我们知道下面的集合都属于 $\sigma(\mathcal{A})$:

(i) $[a, +\infty) = (-\infty, a)^c$;

(ii) $[a,b) = (-\infty, b) \cap [a, +\infty)\ (b > a)$;

(iii) $\{a\} = \bigcap_{i=1}^{\infty} \left[a, a + \frac{1}{i}\right)$;

(iv) $(a,b) = [a,b) \cap a^c$。

在问题 A.1,读者被要求证明所有有理数和所有无理数的集合都是 Borel 集。因此,在$\mathbb{R}$ 中所有熟悉的子集都是 Borel 集。关于非 Borel 集的一个例子,参见 Royden(1988, pp. 64—66)。对于应用研究者,可以忽略相当少见的非

Borel 集的情况，因为在应用研究中，(几乎)从来不会碰到非 Borel 集的情况。

可以在$\mathbb{R}$ 的 Borel 集上定义测度。

定义 A.8(测度) 给定 Ω 子集的类别$\mathcal{F}$，一个测度 μ 是从 Ω 到$\mathbb{R}^+ = \{x \in \mathbb{R} : x \geqslant 0\}$ 的一个映射，满足

(i) $\mu(A) \geqslant 0$，所有 $A \in \mathcal{F}$。

(ii) $\mu(\varnothing) = 0$

(iii) 若$\{A_i\}_{i=1}^{\infty}$是互斥的集合，那么

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty}\mu(A_i)$$

若 $\mu(\Omega) = 1$，那么我们就说 μ 是一个概率测度。

定义 A.9(Lebesgue 测度) Lebesgue 测度是在实际中最有用的测度之一，并且也可能是最直观的。一个 Lebesgue 测度 m 是一个被定义在实轴$\mathbb{R}$（事实上在 Borel 集$\mathcal{B}$)上的一个测度，对于任何区间$[a,b]$$(b \geqslant a)$，具有性质 $m[a,b] = b-a$，即一个区间的 Lebesgue 测度等同于区间的长度。任何单独的点的 Lebesgue 测度为 0。同样，$\mathbb{R}$ 的一个可数子集的 Lebesgue 测度是 0。实轴$\mathbb{R}$ 的 Lebesgue 测度是∞（正如它有无限的长度)。

$\mathbb{R}^2$的一个子集 A 的 Lebesgue 测度给出了 A 面积的一个测度，并且对于一个集合 $A \in \mathbb{R}^3$，它测度了 A 的体积。更高维的 Lebesgue 测度是类似被定义的。 667

然而，注意并非所有的测度可以被解释为测量一个集合的长度，或者面积，或者体积。例如，一个概率测度(定义如下)映射一个具有无限 Lebesgue 测度的集合(比如，无限长度或者体积)到一个在 0 和 1 之间的数字，它当然不是一个集合长度(或者体积)的测度。

定义 A.10(概率作为一个集合函数 P) 一个概率测度是一个从样本空间到单位区间$[0,1]$的一个映射，即 $P:\Omega \to [0,1]$使得

(i) $P(A) \geqslant 0$，所有 $A \in \mathcal{F}$。

(ii) $P(\Omega) = 1$。

(iii) 若$\{A_i\}_{i=1}^{\infty}$是$\mathcal{F}$中互斥的事件，即对所有 $i \neq j$，$A_i \cap A_j = \varnothing$，那么

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$$

三元总体$(\Omega, \mathcal{F}, P)$被称为概率空间。

定义 A.11(几乎处处和空集) 令$\mathcal{F}$是一个定义在 Ω 上的 σ 域，μ 是对所有 $A \in \mathcal{F}$定义的一个测度；那么我们说对任何 $A, B \in \mathcal{F}$，若对于 $A \neq B$ 的元素的集合有一个(μ)测度 0，A 几乎处处等于 B(在 μ 下)。一个零测度集合也被称为一个空集。

例 A.7 考虑两个定义在$[0,1]$上的函数,对所有$x\in[0,1]$,$f(x)=1$;若x是一个无理数,$g(x)=1$;若x是一个有理数,$g(x)=0$。那么对几乎所有x,$f(x)=g(x)$(在 Lebesgue 测度下),因为对$f(x)\neq g(x)$的x的集合是$[0,1]$间有理数的集合,其有一个 Lebesgue 零测度。

在例子 A.7 中,我们引入了在$\mathbb{R}$上的 Borel 集。现在我们定义在$\mathbb{R}^q$上的 Borel 可测集。

定义 A.12(Borel 可测集) 若下面的条件成立,我们说一个集合的集合$\mathcal{A}\in\mathbb{R}^q$是一个 Borel 可测集:

(i) 若$A\in\mathcal{A}$,那么$A^c\in\mathcal{A}$。

668

(ii) 若$A_1,A_2,\cdots\in\mathcal{A}$,那么$\bigcup_{i=1}^{\infty}A_i\in\mathcal{A}$。

$\mathbb{R}^q$的所有熟悉的子集,诸如单元素,开的、闭合的、半开的、半闭的区间,它们的并和/或交,都是 Borel 集。

定义 A.13(Borel 可测函数) 令$f(x)$是一个实值函数,它映射$\mathbb{R}^q:\to\mathbb{R}$ $(x\in\mathbb{R}^q)$。那么$f(\cdot)$被称为一个 Borel 可测函数,若对每一个$a\in\mathbb{R}$,集合$(x\in\mathbb{R}^q:f(x)\leqslant a)$是一个 Borel 可测集。

所有熟悉的函数都是 Borel 可测函数。例如,任何连续函数是 Borel 可测的。同样,一个具有可数的很多非连续点的函数也是 Borel 可测函数。事实上,一个处处不连续的函数也可以是一个 Borel 可测函数(见习题 A.2)。

令X是一个随机变量(从Ω到$\mathbb{R}$的一个映射),我们使用$F(\cdot)$表示X的累积分布函数(CDF),其被定义为$F(x)=P[X\leqslant x]$,对所有$x\in\mathbb{R}$。若$F(\cdot)$是可微的(比如,一个连续随机变量),那么概率密度函数(PDF)被定义为$f(x)=\mathrm{d}F(x)/\mathrm{d}x$。注意也可以对离散随机变量使用 Dirac delta 函数来定义一个概率密度函数;参见定义 A.40 以及后面的讨论。

定义 A.14(Riemann 积分) 令$f(x)$是一个在$[a,b]\subset\mathbb{R}$的连续函数。划分$[a,b]$为$a=x_0<x_1<\cdots<x_n=b$,定义$\Delta x_j=x_j-x_{j-1}$,$j=1,\cdots,n$,并且令$\Delta x=\max_{\{1\leqslant j\leqslant n\}}\Delta x_j$,若极限(A.1)存在,

$$\lim_{\Delta x\to 0}\sum_{j=1}^{n}f(\tilde{x}_j)\Delta x_j \tag{A.1}$$

其中$\tilde{x}_j\in[x_{j-1},x_j]$,那么就被称为$f(\cdot)$在$[a,b]$上的 Riemann 积分并且被表示为

$$\int_a^b f(x)\,\mathrm{d}x\equiv\int_{[a,b]}f(x)\,\mathrm{d}x$$

可以证明，Riemann 积分具有性质

$$\int_a^b f(x)\,\mathrm{d}x = -\int_b^a f(x)\,\mathrm{d}x$$

同样有

$$\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x$$

给定所有的积分都是良好定义且有限的。

Riemann 积分是在微积分教材中经常遇到的积分，并且被应用研究者(物理学家、工程师、经济学家等)所使用。

定义 A.15(Riemann-Stieltjes 积分) 令 $f(x)$ 和 $g(x)$ 是定义在 $[a,b]$ 上的实值有界函数。做一个划分 $a = x_0 < x_1 < \cdots < x_n = b$，定义 $\Delta x_j = x_j - x_{j-1}, j = 1,\cdots,n$，并且令 $\Delta x = \max_{\{1\leq j\leq n\}} \Delta x_j$，若极限(A.2)存在，

669

$$\lim_{\Delta x\to 0}\sum_{j=1}^{n} f(\tilde{x}_j)\left[g(x_{j+1}) - g(x_j)\right] \tag{A.2}$$

其中 $\tilde{x}_j \in [x_{j-1}, x_j]$，那么它就被称为 Riemann Stieltjes 积分并且被表示为 $\int_a^b f(x)\,\mathrm{d}g(x)$。

注意若 $f(\cdot)$ 和 $g(\cdot)$ 有一个共同的非连续点，那么积分不存在。但是若 $f(\cdot)$ 是连续的，$g(\cdot)$ 是可微的，并且 $g^{(1)}(x) = \mathrm{d}g(x)/\mathrm{d}x$ 是 Riemann 积分，那么

$$\int_a^b f(x)\,\mathrm{d}g(x) = \int_a^b f(x)g^{(1)}(x)\,\mathrm{d}x \tag{A.3}$$

式(A.3)的右边是一个 Riemann 积分(即 $m(x) = f(x)g^{(1)}(x)$)。

Riemann-Stieltjes 积分是 Riemann 积分的一般化。当 $g(x) = x$ 时，它回到 Riemann 积分。

定义 A.16(简单函数) 令 $A \subset \mathcal{B}$(Borel 集)，并且令 $f = \mathbf{1}_A$ 表示一个示性函数，使得对 $x \in A$，$f(x) = 1$，否则为 0。我们说 $g(\cdot)$ 是一个简单函数，若它有以下形式：

$$g(x) = \sum_{j=1}^{m} c_j \mathbf{1}_{A_j} \tag{A.4}$$

其中 $A_j \in \mathcal{B}, j = 1,\cdots,m$。

定义 A.17(Lebesgue-Stieltjes 积分) 令 $f(\cdot)$ 是一个(Borel)可测函数，且令 $\mu(A)$ 表示 $A \in \mathcal{B}$ 的 Lebesgue 测度。若 $f = \sum_{j=1}^{m} c_j \mathbf{1}(A_j)$ 是一个简单函数，那么 $f(\cdot)$ 的 Lebesgue-Stieltjes 积分被定义为

$$\int f \mathrm{d}\mu = \sum_{j=1}^{m} c_k \mu(A_j)$$

670 若$f(\cdot)$是一个非负的可测函数,那么 Lebesgue-Stieltjes 积分被定义为

$$\int_A f\mu(\mathrm{d}x) = \sup\left\{\int g\mu(\mathrm{d}x): g \text{ 是一个简单函数}, \text{对所有 } x \in A \text{ 有 } g(x) \leqslant f(x)\right\}$$

在上面的定义中可以将 Lebesgue 测度μ替换为任何其他的测度ν,比如说一个概率测度。这导致

$$\int_A f\nu(\mathrm{d}x) = \sup\left\{\int g\nu(\mathrm{d}x): g \text{ 是一个简单函数}, \text{对所有 } x \in A \text{ 有 } g(x) \leqslant f(x)\right\}$$

我们可以把任何的可测函数$f(\cdot)$表达为$f=f_+ - f_-$,其中$f_+ = \max(f,0)$和$f_- = \max(-f,0)$都是非负的函数,且 Lebesgue-Stieltjes 积分被定义为

$$\int f \mathrm{d}\mu = \int f_+ \mathrm{d}\mu - \int f_- \mathrm{d}\mu$$

注意 Lebesgue-Stieltjes 积分是 Riemann-Stieltjes 积分的一个进一步扩展。即使一个处处不连续的函数(其当然不是 Riemann 积分)也可以是 Lebesgue 可积的(见习题 A.7)。

即使一个 Lebesgue-Stieltjes 积分函数也可能不是 Riemann 可积的。下一个定理告诉我们 Riemann 和 Lebesgue-Stieltjes 积分是紧密相关的。

定理 A.1 假定$f(\cdot)$是一个有界函数,是在A上的 Riemann 积分,我们用$(R)\int_A f$来表示,那么它也是 Lebesgue-Stieltjes 积分且两个积分是相同的,即

$$\int_A f(x)\mu(\mathrm{d}x) = (R)\int_A f(x)\,\mathrm{d}x$$

其中μ是 Lebesgue 测度。

证明:参见 Wheeden 和 Zygmund(1977,定理 5.52)。 □

定理 A.1 阐述了所有的 Riemann 积分函数也是 Lebesgue-Stieltjes 积分。此671 外,两个积分的结果是相同的。考虑到一个 Riemann 积分函数总体上比一个 Lebesgue-Stieltjes 积分更容易计算,人们总是可以使用前者来获得后者的结果,只要 Riemann 积分存在(参见习题 A.8)。然而,也存在一些函数不是 Riemann 积分,而是 Lebesgue-Stieltjes 积分。例如,Dirichlet 函数就是这样一个例子(见习题 A.2)。

一个随机变量X的期望被定义为

$$E(X) = \int xF(\mathrm{d}x) \equiv \int x\mathrm{d}F(x)$$

其中 $F(\cdot)$是 X 的累积分布函数。它被定义为一个 Lebesgue-Stieltjes 积分。然而,若 X 的概率密度函数 $f(x)$存在,那么

$$E(X) = \int xf(x)\mathrm{d}x$$

这个积分变成一个 Riemann 积分,其通常比一个 Lebesgue-Stieltjes 积分更容易计算。

定理 A.2 令 X 有一个连续的累积分布函数 $F_X(x)$,并且令 $Y = F_X(X)$。那么 Y 在$[0,1]$上均匀分布,即对于任何 $y \in [0,1]$,$P(Y \leqslant y) = y$。

证明:

$$\begin{aligned} P(Y \leqslant y) &= P[F_X(x) \leqslant y] \\ &= P\{F_X^{-1}[F_X(x)] \leqslant F^{-1}(y)\} \quad (F^{-1}X \text{ 是递增的}) \\ &= P(X \leqslant F_X^{-1}(y)) \quad (\text{见如下的论证}) \\ &= F_X(F_X^{-1}(y)) \quad (F_X(\cdot) \text{ 的定义}) \\ &= y \quad (F_X(\cdot) \text{ 的连续性}) \end{aligned}$$

在证明中,我们使用了

$$P\{F_X^{-1}[F_X(X)] \leqslant F^{-1}(y)\} = \mathrm{P}(X \leqslant F_X^{-1}(y)$$

当 $F_X(\cdot)$是严格递增时它是真的。然而,$F_X(\cdot)$可能是平坦的;当 $F_X(\cdot)$平坦时,具体参见 Casella 和 Berger(2002, p. 55)。

定义 A.18(特征函数) 令 X 是一个在$\mathbb{R}^q$ 上具有累积分布函数 $F(\cdot)$的随机向量。X 的特征函数是一个复杂的值函数,被定义为

$$\phi(t) = E(\mathrm{e}^{it'X}) = \int \mathrm{e}^{it'x}\mathrm{d}F(x) \tag{A.5}$$

其中 $i = \sqrt{-1}$,$t'x = t_1x_1 + t_2x_2 + \cdots + t_qx_q$。积分总是存在,因为

$$|\mathrm{e}^{it'x}| = |\cos(it'x) + i\sin(it'x)| = \sqrt{\cos^2(it'x) + \sin^2(it'x)} = 1$$

对于一个连续随机变量,$\phi(\cdot)$唯一确定它的概率密度函数为下面的反算公式 672 (inversion formula) $\left(\text{给定}\int|\phi(t)|\,\mathrm{d}t < \infty\right)$:

$$f(x) = \frac{1}{(2\pi)^q}\int \mathrm{e}^{-it'x}\phi(t)\mathrm{d}t \tag{A.6}$$

证明:从式(A.5)我们知道

$$\phi(t) = \int f(v)\mathrm{e}^{it'v}\mathrm{d}v$$

将其代入式(A.6)中,我们得到(也改变了双重积分的顺序)

$$\int e^{-it'x}\phi(t)\,dt = \frac{1}{2\pi}\int\left[\int e^{-it'(x-v)}\,dt\right]f(v)\,dv$$
$$= \int \delta(x-v)f(v)\,dv = f(x)$$

其中我们利用了以下事实:

$$\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it'v}\,dt = \delta(v)$$

即 Dirac delta 函数,我们将在下面的定义 A.40 和方程(A.15)中定义它。

定义 A.19(Lipschitz 函数) 令 $g(\cdot)$是一个$\mathbb{R}^q$上的实值函数。它是阶数为 1 的一个 Lipschitz 函数,若如下的不等式被满足:

$$|g(x)-g(y)| \leqslant c\|x-y\|$$

对所有 $x,y\in\mathbb{R}^q$,其中 c 是一个有限的常数。

1.2 度量、范数和泛函空间

定义 A.20(线性向量空间) 一个元素的集合$\mathcal{V}$被称为实数上的一个向量空间(或者线性空间,或者线性向量空间),若我们在$\mathcal{V}\times\mathcal{V}$到$\mathcal{V}$上定义一个 +(加法)以及在$\mathbb{R}\times\mathcal{V}$到$\mathcal{V}$上定义一个 ·(标量乘法)满足以下的条件:

(i) $x+y=y+x$。

(ii) $(x+y)+z=x+(y+z)$。

(iii) 有一个向量 θ 使得 $x+\theta=x$,对所有 $x\in\mathcal{V}$。

(iv) $\alpha(x+y)=\alpha x+\alpha y$,对所有 $\alpha\in\mathbb{R}$,且所有 $x,y\in\mathcal{V}$。

(v) $(\alpha+\beta)x=\alpha x+\beta x$,对所有 $\alpha,\beta\in\mathbb{R}$,且所有 $x\in\mathcal{V}$。

673

(vi) $\alpha(\beta)x=(\alpha\beta)x$,对所有 $\alpha,\beta\in\mathbb{R}$,且所有 $x\in\mathcal{V}$。

(viii) $0\cdot x=\theta$, $1\cdot x=x$。

以上定义的 θ 元素可以被证明是唯一的,且其被称为零元素。

定义 A.21($C[a,b]$和 $C^m[a,b]$空间) 我们使用 $C[a,b]$来表示在$[a,b]$上有界且连续的实值函数的集合。类似地,我们使用 $C^m[a,b]$来表示在$[a,b]$上 m 次连续可微的函数(具有有界的导数),其中 m 是一个非负的整数。

定义 A.22($L^p[a,b]$空间) 令 p 是一个正的实数。一个定义在$[a,b]$上的可测函数被称为是属于空间 $L^p=L^p[a,b]$的,若$\int_a^b|f(x)|^p dx<\infty$。

根据加法和标量乘法的通常定义,$C^m[a,b]$和 $L^p[a,b]$都是线性向量空间。

定义 A.23(线性空间)　一个实值函数的空间 X 被称为一个线性空间,如果它有以下性质:若 f 和 $g \in X$,那么 $\alpha f+\beta g \in X$,其中 α 和 β 是任意常数。例如,L^p 空间是一个线性空间。

定义 A.24(扩张空间和生成基)　令 $v_1,\cdots,v_n$ 是一个向量空间 V 中的向量,一个形如 $\alpha_1 v_1+\alpha_2 v_2+\cdots+\alpha_n v_n$ 的加总被称为 $v_1,\cdots,v_n$ 的一个线性组合,其中 $\alpha_1,\cdots,\alpha_n$ 是标量。$v_1,\cdots,v_n$ 的所有线性组合的集合将被表示为 $\mathrm{Span}(v_1,\cdots,v_n)$。例如,令 $v_1=(1,0)',v_2=(0,1)$ 是 $\mathbb{R}^2(V=\mathbb{R}^2)$ 中的两个向量,那么 $\mathrm{Span}(v_1,v_2)=\mathbb{R}^2$。也就是说,两个向量 v_1 和 v_2 扩张空间 $\mathbb{R}^2$,(v_1,v_2) 被称为 $\mathbb{R}^2$ 的生成基。

定义 A.25(一个赋范的线性空间)　一个线性空间被称为是一个赋范的线性空间,若我们分配一个非负的实数(一个范数)$\|f\|$ 到每一个 $f\in X$ 使得

(i) $\|\alpha f\|=|\alpha|\,\|f\|$。

(ii) $\|f+g\|\leqslant\|f\|+\|g\|$。

(iii) $\|f\|=0$,当且仅当 $f\equiv 0$。

L^p 空间是一个赋范线性空间,其范数被定义为

674

$$\|f\|=\left\{\int|f(x)|^p\mathrm{d}x\right\}^{1/p}$$

令 $\|\cdot\|_{\mathcal{V}}$ 是一个向量空间 $\mathcal{V}$ 上的一个范数。一个序列 $\{v_k\}_{k=1}^{\infty}$ 被称为一个 Cauchy 序列,若当 $j,k\to\infty$ 时,$\|v_j-v_k\|_{\mathcal{V}}\to 0$。

定义 A.26(Banach 空间)　一个赋范线性空间被称为完全的,若在空间里的每一个 Cauchy 序列收敛。一个完全的赋范线性空间被称为一个 Banach 空间。

Banach 空间的例子包括 $C[a,b]$,具有以 $\|f-g\|_{\sup}=\sup_{x\in[a,b]}|g(x)-f(x)|$ 定义的范数;或者 $L^2[a,b]$,具有以 $\|g-f\|_{L_2}=\left\{\int_a^b[g(x)-f(x)]^2\mathrm{d}x\right\}^{1/2}$ 定义的范数。

一个内积是一个被定义在 $V\times V:\to\mathbb{R}$ 上的函数,具有以下性质:对所有 $x,y,z\in V$ 和所有 $\alpha,\beta\in\mathbb{R}$,

(i) $(\alpha x+\beta y,z)=\alpha(x,z)+\beta(y,z)$。

(ii) $(x,y)=(y,x)$。

(iii) $(x,x)=\|x\|^2$。

定义 A.27(Hilbert 空间)　一个 Hilbert 空间是一个其范数通过一个内积来定义的 Banach 空间。

例 A.8　例如,$L^2[0,1]$ 是一个 Hilbert 空间,具有以 $(g,f)=\int f(x)g(x)\mathrm{d}x$

定义的内积，以及 L_2-范数 $\| f-g \|_{L_2} = \{(f-g,f-g)\}^{1/2}$。然而，其范数以 $\| g-f \|_{\sup} = \sup_{x\in[a,b]} |g(x)-f(x)|$ 定义的 $C[a,b]$ 不是一个 Hilbert 空间，因为不存在一个内积能够乘以 sup 范数 $\| \cdot \|_{\sup}$。

定义 A.28（正交基） 令 $\mathcal{H}$ 表示一个 Hilbert 空间，令 $e_1, e_2, \cdots$，是 $\mathcal{H}$ 中的元素，且令 $(\cdot,\cdot)$ 表示内积。我们说 $\{e_j\}_{j=1}^{\infty}$ 是 $\mathcal{H}$ 的一个完全（可数）基，若对所有 $g\in\mathcal{H}$，我们有

$$g = \sum_{j=1}^{\infty} c_j e_j \tag{A.7}$$

其中 c_j 是某些常数。

若 $(e_i,e_j)=\delta_{ij}$ 对所有 $i,j\in\mathbb{N}^+$（$\delta_{ij}=1$，若 $i=j$；否则为 0），我们说 $\{e_i\}_{i=1}^{\infty}$ 是一个正交基。

若 $\{e_j\}$ 是一个正交基，那么很容易证明在式（A.7）中 $c_j=(e_j,g)$。

675 注意我们仅仅讨论 Hilbert 空间有一个可数（完全的）基的情况。一般而言，一个 Hilbert 空间可能有一个不可数的基。然而，在多数应用中，我们仅仅遇到前者的情况。

定义 A.29（Parseval 等式） 令 $\{e_j\}_{j=1}^{\infty}$ 是 $\mathcal{H}$ 的一个完全基（一个 Hilbert 空间），那么对所有 $g\in\mathcal{H}$，我们有

$$\| g \|^2 \overset{\text{def}}{=} (g,g) = \sum_{j=1}^{\infty} (g,e_j)^2 \tag{A.8}$$

例如，令 $L^2[-\pi,\pi]$ 表示有界连续和 $[-\pi,\pi]$ 上的平方积分函数，那么序列 $\{1/\sqrt{2\pi}, \sin(\pi x)/\pi, \cos(\pi x)/\pi, \sin(2\pi x)/\pi, \cos(2\pi x)/\pi, \cdots\}$ 形成一个对于 $\mathcal{H}$ 的完全正交基（见习题 A.10）。

另一个在 $L^2[-1,1]$ 上的正交基的例子是 Legendre 多项式 $P_n(x)$（即在 $[-1,1]$ 上的一个正交基上，其被定义为

$$P_n(x) = \frac{(2n+1)^{1/2}}{2^{n+1/2}n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2-1)^n,\ n=0,1,2,\cdots \tag{A.9}$$

例 A.9 $P_0(x)=1/2$，$P_1(x)=\sqrt{3/2}x$，$P_2(x)=(3/2)\sqrt{5/2}(x^2-1/3)$，等等。可以证明，对 $j\neq l$

$$\int_{-1}^{1} P_j(x)P_l(x) = 0$$

同样也可以证明

$$\int_{-1}^{1} P_n{}^2(x)\,\mathrm{d}x = 1$$

使得 $\{P_j(x)\}_{j=0}^{\infty}$ 形成对 $L^1[-1,1]$ 的一个正交基。

定义 A.30(Bessel 不等式) 令$\{v_j\}_{j=1}^{\infty}$是$\mathcal{H}$(一个 Hilbert 空间)上的一个正交序列。那么对任何$g \in \mathcal{H}$,我们有

$$\sum_{j=1}^{\infty} (g, v_j)^2 \leqslant \| g \|^2 \tag{A.10}$$

定义 A.31(Sobolev 范数) 对于定义在$[a,b]$上的可微函数,(一阶)Sobolev 范数被定义为

$$\| f \|_H = \left\{ \int_a^b f(x)^2 \mathrm{d}x + \int_a^b \left(\frac{\mathrm{d}f}{\mathrm{d}x} \right)^2 \mathrm{d}x \right\}^{1/2}$$

对于p阶可微的多元函数$(f(x), x \in \mathbb{R}^q)$, Sobolev 范数被定义为

676

$$\| f \|_{H^p} = \left\{ \sum_{0 \leqslant p_1 + \cdots + p_q \leqslant p} \int_a^b \left[\frac{\partial f(x)}{\partial x_1^{p_1} \cdots \partial x_q^{p_q}} \right]^2 \mathrm{d}x \right\}^{1/2}$$

定义 A.32(度量空间) 一个度量空间$\{X, \rho\}$是一个非空集合X的元素和定义在$X \times X$上的一个实值函数ρ的结合,使得对于所有在X中的x、y和z,有

(i) $\rho(x,y) \geqslant 0$。

(ii) $\rho(x,y) = 0$,当且仅当$x = y$。

(iii) $\rho(x,y) = \rho(y,x)$。

(iv) $\rho(x,y) \leqslant \rho(x,z) + \rho(z,y)$。

函数$\rho(\cdot,\cdot)$被称为一个度量(metric)。

度量的概念是距离的一般化。一个简单的度量空间的例子是具有

$$\rho(x,y) = \| x - y \| = \sqrt{x_1^2 + \cdots + x_q^2}$$

($\| \cdot \|$是欧几里得范数)的所有数的集合$\mathbb{R}^q$。事实上,对于任何赋范空间(具有范数$\| \cdot \|$),可以通过$\rho(x,y) = \| x - y \|$来定义一个度量,这个度量被称为通过范数$\| \cdot \|$导致的。当在一个度量空间的定义中去掉条件(ii),即当我们允许对某些$x \neq y$有$\rho(x,y) = 0$的可能性时,ρ被称为一个伪度量(psendometric)。例如,空间L^p具有以$\rho(f,g) = \{\int |f(x) - g(x)|^p \mathrm{d}x\}^{1/p}$定义的度量,是一个定义在$L^p$上的伪度量。然而,若我们把$f = g$ a.e. 处理成$f = g$,那么L^p变成一个度量空间。

定义 A.33(开集) 若对每一个$x \in A$,有一个$\delta > 0$,使得每一个y有$\rho(y, x) < \delta$都属于A,在度量空间$\{X, \rho\}$中的一个集合B被称为开的(一个开集)。

定义 A.34(闭包) 若对每一个$\delta > 0$有一个$y \in B$,使得$\rho(y,x) < \delta$,元素x被称为一个集合B闭包的一个点。我们使用$\overline{B}$来表示B的闭包。很明显,$B \subset \overline{B}$。

定义 A.35(闭集) 若$B = \overline{B}$,一个在$\{X, \rho\} f$的集合B被称为是闭的(一个闭集)。

677 例如,对于具有 $\rho(x,y)=|x-y|$ 的实轴 $\mathbb{R}$,我们有 $A=(0,1)$ 是一个开集,A 的闭包 $\bar{A}=[0,1]$ 是一个闭集。

定义 A.36(稠密) 令 B 是在度量空间上(具有度量 ρ)的一个闭集。若对所有(小的)$\varepsilon>0$,且对每一个 $x\in B$,有一个 $y\in A$,使得 $\rho(x,y)<\varepsilon$,集合 $A\subset B$ 被称为在 B 上的稠密。①

例如,有理数的集合在实数集合(具有 $\rho(x,y)=|x-y|$)中是稠密的。同样,$C^1[a,b]$ 在具有以 L_2 范数(度量)导致的度量 $C[a,b]$ 中是稠密的,即

$$\rho(f,g)=\left\{\int_a^b[f(x)-g(x)]^2\mathrm{d}x\right\}^{1/2}$$

定义 A.37(绝对连续) 若对于任何集合 $A\in\Omega,\mu(A)=0$ 意味着 $\nu(A)=0$,则测度 ν 被称为关于测度 μ 是绝对连续的。

令 $f(\cdot)$ 是一个非负的 Borel 可测函数,可以通过下式定义集合函数:

$$\lambda(A)=\int_A f\mathrm{d}\nu,\quad A\in\mathcal{B}$$

其是 $(\Omega,\mathcal{F})$ 上的一个测度。可以证明 $\nu(A)=0$ 意味着 $\lambda(A)=0$。那么,λ 关于 ν 是绝对连续的。$f(\cdot)$ 被称为 λ 关于 ν 的 Radon-Nikodym 导数或者密度且被表示为 $f=\mathrm{d}\lambda/\mathrm{d}\nu$。

例 A.10 令 $f(\cdot)$ 是一个概率密度函数,那么对应的累积分布函数被定义为

$$F(x)=\int_{(-\infty,x]}f(u)\mathrm{d}u$$

我们知道 $\mathrm{d}F(x)/\mathrm{d}x=f(x)$,即 $f(\cdot)$ 是 $F(\cdot)$ 的 Radon-Nikodym 导数或者密度,而 $F(\cdot)$ 是一个概率测度。此处 $\mathrm{d}x=\mathrm{d}\nu$,其中 ν 是 Lebesgue 测度,且若 $f(\cdot)$ 是一个可测的(比如,连续的)函数,那么 $F(\cdot)$ 是关于 Lebesgue 测度的绝对连续。

注意,若一个累积分布函数 $F(x)$ 有非连续的点,如离散随机变量的情况,那么它关于 Lebesgue 测度不是绝对连续。

绝对连续比可微更弱,但强于连续(对 $F(\cdot)$)。

678 **定义 A.38(标准布朗运动(或者维纳过程))** 我们称一个定义在 $[0,1]$ 上的 q 维随机过程 $W(t)$ 为一个 q 维的标准布朗运动,若

(i) $P[W(0)=0]=1$。

(ii) $W(t_1)-W(t_0),W(t_2)-W(t_1),\cdots,W(t_n)-W(t_{n-1})$ 独立于任何正整

① 通常,"A 在 B 中稠密"被定义为 A 的闭包等同于没有使用距离(度量)概念的 B;参见 Royden (1988, p.142)。此处我们在度量空间中给出一个"稠密"的等价定义。

数 n 和时间划分 $0 \leqslant t_0 < t_1 < \cdots < t_n \leqslant 1$。

(iii) $W(t) - W(s) \sim N(0, (t-s) I_q)$, $0 \leqslant s < t \leqslant 1$。

定义 A.39(高斯过程)　一个随机过程 $Z(\cdot)$(或者 $Z(x)$,以 x 标记,其中 x 属于$\mathbb{R}^q$中的一个紧集)被称为一个高斯过程,若每一个它的有限维边缘$(Z(x_1), \cdots, Z(x_m))$在欧几里得空间$\mathbb{R}^m$上有一个多元的正态分布。

注意 $W(t)$是具有独立增量和协方差结构 $\mathrm{cov}(W(s), W(t)) = \min(s,t) I_q$ 的一个零均值非平稳高斯过程。可知 $W(t)$的样本路径以概率 1 连续,而它在$[0,1]$上的任何区间子集是无处可微的。

令 $X(t)$是一个满足如下(Ito)随机微分方程的随机过程:

$$\mathrm{d}X(t) = \mu(X(t), t)\mathrm{d}t + \sigma(X(t), t)\mathrm{d}W(t), \quad 0 \leqslant t \leqslant 1 \tag{A.11}$$

引理 A.1(Ito 引理)　令 $g(x,t)$是一个$(-\infty, \infty) \times [0,1]$上的一个连续函数(且假定一些附加的正则性条件成立;参见 Tanaka(1996, p. 58)),那么 $g(x,t)$满足随机微分方程($X_t \equiv X(t)$)

$$\mathrm{d}g(X_t, t) = \frac{\partial g}{\partial X_t}\mathrm{d}X_t + \left(\frac{\partial g}{\partial t} + \frac{1}{2}\frac{\partial g^2}{\partial X_t^2}\sigma^2(X_t, t)\right)\mathrm{d}t \tag{A.12}$$

将式(A.12)中的 $\mathrm{d}X_t$ 替换为式(A.11),我们得到

$$\mathrm{d}g(X_t, t) = \left[\frac{\partial g}{\partial t} + \frac{\partial g}{\partial X_t}\mu(X_t, t) + \frac{1}{2}\frac{\partial^2 g}{\partial X_t^2}\sigma^2(X_t, t)\right]\mathrm{d}t + \frac{\partial g}{\partial X_t}\sigma(X_t, t)\mathrm{d}W_t \tag{A.13}$$

现在,若 $g(x,t) = \mu(x)$ 以及 $\sigma^2(x,t) = \sigma^2(x)$ 都是时间同质过程,那么式(A.13)简化为　679

$$\mathrm{d}\mu(x) = \mu(x)\mu^{(1)}(x)\mathrm{d}t + \frac{1}{2}\mu^{(2)}(x)\sigma^2(x)\mathrm{d}t + \mu^{(1)}(x)\sigma(x)\mathrm{d}W_t \tag{A.14}$$

定义 A.40(Dirac delta 函数)　Dirac delta 函数通过 $\delta(x)$来表示,且有如下性质($x \in \mathbb{R}^q$):

(i) $\delta(x) = 0, x \neq 0$。

(ii) $\delta(0) = \infty$。

(iii) $\int_{-\infty}^{\infty} \delta(x)\mathrm{d}x = 1$。

可以证明对任何可测函数 $g(x)$,我们有

$$\int_{-\infty}^{\infty} \delta(x) g(x)\mathrm{d}x = g(0)$$

类似地,对任何实数 a,我们有(见习题 A.11)

$$\int_{-\infty}^{\infty}\delta(x-a)g(x)\mathrm{d}x = g(a)$$

对 Dirac delta 函数有许多(等价的)表达式。可以证明($t,x\in\mathbb{R}^q$),

$$\delta(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{e}^{it'x}\mathrm{d}x \tag{A.15}$$

也可以把 Dirac delta 函数定义为一个阶段函数的导数。考虑阶段函数 $\mathbf{1}(x\geqslant 0)$,当 $x<0$ 时取 0,$x\geqslant 0$ 时取 1。这个函数在 0 处是不连续的;因此,在通常意义上,它在 $x=0$ 是不可微的。但是,对所有 $x\neq 0$,可以说它有导数 0,且它在 $x=0$ 有一个无穷大的导数,这意味着 $\mathbf{1}(x\geqslant 0)$ 的导数是一个 Dirac delta 函数,即 $\mathrm{d}\mathbf{1}(x\geqslant 0)/\mathrm{d}x=\delta(x)$。

因此,人们可以使用 Dirac delta 函数来对离散随机变量定义一个概率密度函数。考虑以 $F_n(x) = n^{-1}\sum_{i=1}^{n}\mathbf{1}(X_i\leqslant x)$ 来定义的经验累积分布函数的情况。对应的经验概率密度函数可以被定义为

$$f_n(x) = \mathrm{d}F_n(x)/\mathrm{d}x = n^{-1}\sum_{i=1}^{n}\delta(X_i - x)$$

那么,我们有

$$\int_{-\infty}^{x}f_n(v)\mathrm{d}v = n^{-1}\sum_{i=1}^{n}\int_{-\infty}^{x}\delta(X_i - v)\mathrm{d}v$$

$$= n^{-1}\sum_{i=1}^{n}\mathbf{1}(X_i\leqslant x)$$

680 其正好是 $F_n(x)$。利用这个经验概率密度函数,很容易证明对任何可测函数 $g(x)$,我们有

$$\int_{-\infty}^{\infty}g(x)\mathrm{d}F_n(x) = \int_{-\infty}^{\infty}g(x)f_n(x)\mathrm{d}x$$

$$= n^{-1}\sum_{i=1}^{n}\int g(x)\delta(X_i - x)\mathrm{d}x = n^{-1}\sum_{i=1}^{n}g(X_i)$$

1.3 极限和收敛的方式

1.3.1 极限上确界和下确界

定义 A.41(上界和下界) 若对所有 $x\in A$,有 $b\geqslant x$,则数 b 被称为是集合 A 的上界。若 c 是 A 的一个上界,且对 A 的每一个上界 b 有 $c\leqslant b$,则数 c 被称为集合 A 的最小上界(用 $\sup A$ 表示)。进一步,若最大元素存在,$\sup A$ 是 A 的最大元素;例如,若 $A=[0,1]$,那么 $\sup A=\max A=1$。但是,对于 $B=(0,1)$, $\sup B$

$=1$,而 $\max B$ 不存在,因为 $1\notin B$("sup"的意思是上确界)。

集合 A 的最大下界 $\inf A$ 是类似被定义的。对于上面 $A=[0,1]$ 的例子,$\inf A=\min A=0$,对 $B=(0,1)$,$\inf B=0$,而 $\min B$ 不存在。

定义 A.42(极限上确界和极限下确界)　令 $\{a_n\}_{n=1}^{\infty}$ 是一个实数序列。对任何正整数 k,令 $\beta_k=\sup\{a_k,a_{k+1},\cdots\}$,其明显是非递增的,且令 $\alpha_k=\inf\{a_k,a_{k+1},\cdots\}$,那么 $\{\alpha_k\}$ 是一个非递减的序列。若 a_n 是一个有界的序列,$\lim\limits_{k\to\infty}\beta_k$ 和 $\lim\limits_{k\to\infty}\alpha_k$ 都存在(有界单调序列肯定收敛)。它们分别被称为 a_n 的极限上确界和极限下确界,且分别被表示为

$$\overline{\lim_{n\to\infty}}\,a_n\equiv\limsup a_n\overset{\text{def}}{=}\lim_{k\to\infty}\sup_{n\geqslant k}a_n$$

$$\underline{\lim_{n\to\infty}}\,a_n\equiv\liminf a_n\overset{\text{def}}{=}\lim_{k\to\infty}\inf_{n\geqslant k}a_n$$

可以证明 $\limsup a_n=\inf\limits_{n}\sup\limits_{k\geqslant n}a_k$,这比 $\{a_n\}$ 的最小上界更直观。类似地,

681

$$\liminf a_n=\sup_{n}\inf_{k\geqslant n}a_k$$

显然,

$$\limsup a_n\geqslant\liminf a_n$$

并且 a_n 的极限存在,当且仅当

$$\limsup a_n=\liminf a_n\left(=\lim_{n\to\infty}a_n\right)$$

例如,若 $a_n=(-1)^n=\{-1,1,-1,1,\cdots\}$,那么 $\limsup a_n=1$ 且 $\liminf a_n=-1$,且 $\lim\limits_{n\to\infty}a_n$ 不存在。

若 $\lim\limits_{n\to\infty}a_n$ 不存在,那么 $\{a_n\}_{n=1}^{\infty}$ 通常有几个收敛子序列。若它们收敛于不同的值(如上面例子表明的),那么 $\limsup a_n$ 等于其中最大的那个,$\liminf a_n$ 等于最小的。在极限中当 $n\to\infty$ 时,或者对 n 足够大的值,a_n 的取值必然在这两个极限值之间。

对于一个一般的集合 $\{A_n\}_{n=1}^{\infty}$,$\limsup A_n$ 被定义如下:$x\in\limsup A_n$,当且仅当对无限多的 n,$x\in A_n$。$\liminf A_n$ 被定义为 $x\in\liminf A_n$,当且仅当对所有的 n,除了可能有限多的例外,$x\in A_n$,即存在一个固定的正整数 n_0,使得 $x\in A_n$,对所有 $n\geqslant n_0$ 时。很显然,$\liminf A_n\subset\limsup A_n$。

可以证明

$$\limsup A_n=\bigcap_{n=1}^{\infty}\bigcup_{k=n}^{\infty}A_k$$

$$\liminf A_n=\bigcup_{n=1}^{\infty}\bigcap_{k=n}^{\infty}A_k$$

总之,$x\in\limsup A_n$,当且仅当对所有 $n=1,2\cdots$,存在一个 $k\geqslant n$,使得 $x\in A_k$。类

似地，$x \in \liminf A_n$，当且仅当对所有 $k \geq n$，存在一个正整数 n，使得 $x \in A_k$。

例如，定义 $A_n = [0,2]$，$n = 1,3,5,\cdots$，以及 $A_n = [0,1]$，$n = 2,4,6,\cdots$，那么，$\limsup A_n = [0,2]$，$\liminf A_n = [0,1]$。

1.3.2 收敛的方式

为了定义各种收敛的方式，我们首先引入一个向量“欧几里得长度”（“欧几里得范数”）的概念。给定一个 $q \times 1$ 的向量 $x = (x_1, \cdots, x_q)' \in \mathbb{R}^q$，我们用 $\| x \|$ 来表示 x 的欧几里得长度，定义为

$$\| x \| = [x'x]^{1/2} \equiv \sqrt{x_1^2 + x_2^2 + \cdots + x_q^2}$$

其中 $q = 1$（一个标量），$\| x \|$ 是 x 的绝对值。

定义 A.43（依概率收敛） 令 $\{\mathcal{X}_n\}_{n=1}^{\infty}$ 是实数随机变量的序列（可能是一个有限维的向量或者矩阵值），令 X 是一个与 $\mathcal{X}_n$ 有相同维数的随机向量。我们说 $\mathcal{X}_n$ 依概率收敛于 X，若对每一个（小的）$\varepsilon > 0$，

682

$$\lim_{n \to \infty} \mathrm{P}(| \mathcal{X}_n - X | < \varepsilon) = 1$$

我们用 $\mathcal{X}_n \xrightarrow{P} X$ 来表示 $\mathcal{X}_n$ 依概率收敛于 X。

定义 A.44（r 次均值收敛） 我们说 $\mathcal{X}_n$ 依 r 次均值收敛于 X，若对某些 $r > 0$，

$$\lim_{n \to \infty} E(\| \mathcal{X}_n - X \|^r) = 0$$

我们用 $\mathcal{X}_n \xrightarrow{r\text{th}} X$ 来表示依 r 次均值收敛。

当 $r = 2$ 时，我们称 $\mathcal{X}_n$ 依均方误差收敛于 X。

定义 A.45（依分布收敛） 我们称 $\mathcal{X}_n$ 依分布收敛于 X，表示为 $\mathcal{X}_n \xrightarrow{d} X$，若对 $F(x)$ 所有的连续点，有

$$\lim_{n \to \infty} F_n(x) = F(x)$$

其中 $F_n(x)$ 和 $F(x)$ 分别是 $\mathcal{X}_n$ 和 X 的累积分布函数。

定义 A.46（几乎确定收敛） 我们说 $\mathcal{X}_n$ 几乎确定（a. s.）收敛于 X（或者以概率 1，几乎处处（a. e.），或者更强地），若

$$P(\lim_{n \to \infty} \mathcal{X}_n = X) = 1$$

（参见 Serfling（1980，p. 6））。

在本书中，尽管我们经常使用依概率收敛的概念，但要指出我们很少使用定义 A.43 来判断一个随机变量的概率极限。这是因为计算一个随机变量均方误差的极限（或者一般而言 r 次均值）比计算随机变量的概率极限容易得多，给定下文的定理 A.3，依均方误差收敛（或者 r 次均值）隐含着依概率收敛。因

此，在实践中，当判断一个随机变量的概率极限时，我们采用最简单的方法，依 r 次均值收敛。

683

定理 A.3　若$\mathcal{X}_n \xrightarrow{rth} X$（依 r 次均值），那么$\mathcal{X}_n \xrightarrow{P} X$。

证明：这要使用 Chebychev 不等式。对任何 $\varepsilon > 0$，通过假定 $X_n \to X$ 依 r 次均值，我们有

$$P(\| X_n - X \| < \varepsilon) \leqslant \frac{1}{\varepsilon^r} E[\| X_n - X \|^r] \to 0$$ □

我们强调定理 A.3 将被经常使用，在本书中将仅仅涉及 $r=1$ 和 $r=2$ 的情形。下面我们提供一个例子来说明通过利用定理 A.3，概率极限的计算事实上可以相当直接。

例 A.11

(i) 令 Y_n 是一个零均值和有限方差独立同分布随机变量的序列（比如独立同分布 $N(0,1)$）定义$\mathcal{X}_n = Y_n/n$。找出均方误差和$\mathcal{X}_n$的概率极限。

证明：当 $n \to \infty$ 时，$E(\mathcal{X}_n^2) = E(Y_n^2)/n^2 = 1/n^2 \to 0$。因此$\mathcal{X}_n \to 0$ 依均方误差收敛。根据定理 A.3（$r=2$），我们知道$\mathcal{X}_n \xrightarrow{p} 0$。

(ii) 令 $X_1, \cdots, X_n$ 是独立同分布．且具有均值 μ 和有限方差 σ^2。找出 $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ 的概率极限。

证明：首先，从直觉意义上可以推测极限是总体均值 μ。我们通过计算下式（当 $n \to \infty$ 时）来证实这一点：

$$E[(\overline{X}_n - \mu)^2] = 1/n^2 \sum_{i=1}^{n}\sum_{j=1}^{n} E[(X_i - \mu)(X_j - \mu)]$$

$$= 1/n^2 \left\{ \sum_{i=1}^{n} E[(X_i - \mu)^2] + 0 \right\} = \sigma^2/n \to 0$$

其中我们利用了 $\mathrm{cov}(X_i, X_j) = 0$ 的事实，因为 X_i 和 X_j 对于 $i \neq j$ 是相互独立的。于是 $\overline{X}_n \xrightarrow{MSE} \mu$，且根据定理 A.3 有 $\overline{X}_n \xrightarrow{p} \mu$。 □

定理 A.4

(i) 若 $X_n \xrightarrow{d} X$，且 $g(\cdot)$ 是一个有界连续函数，那么 $\int g\mathrm{d}F_n \to \int g\mathrm{d}F$ 依分布收敛。

(ii) 若 $X_n \xrightarrow{d} X, Y_n \xrightarrow{p} c, g(\cdot)$ 是一个连续函数，那么 $g(X_n, Y_n) \xrightarrow{d} g(X, c)$。

接下来的定理提供了依概率收敛和依分布收敛之间的关系。

684

定理 A.5　若$\mathcal{X}_n \xrightarrow{p} X$，那么$\mathcal{X}_n \xrightarrow{d} X$。

证明：参见 Serfling(1980, p. 19)。 □

注意定理 A.3 和 A.5 的逆命题可能不成立(参见 Serfling(1980)相应的例子)。

在下面的定理中我们提供了一些关于依概率收敛和依分布收敛的有用的性质。这些结论在本书中经常被用到。

定理 A.6 若$\mathcal{X}_n \xrightarrow{d} X$且$\mathcal{Y}_n \xrightarrow{p} c$,其中 c 是一个常数,那么

(i) $\mathcal{X}_n + \mathcal{Y}_n \xrightarrow{d} X + c$。

(ii) $\mathcal{X}_n \mathcal{Y}_n \xrightarrow{d} cX$。

(iii) $\mathcal{X}_n / \mathcal{Y}_n \xrightarrow{d} X/c$(给定 $c \neq 0$)。

证明:参见 Serfling(1980, p. 19)。 □

上面的性质类似于普通的极限性质。在一些更强条件下(i) 的证明被留作习题(参见习题 A.4)。

定义 A.47(阶数:大 $O(\cdot)$和小 $o(\cdot)$) 对于一个正整数 n,我们记 $a_n = O(1)$,若当 $n \to \infty$ 时,a_n 仍然是有界的,即对某些常数 C 以及对所有 n 的很大的值(a_n 是一个有界序列),有$|a_n| \leqslant C$。

我们记 $a_n = o(1)$,若当 $n \to \infty$ 时,$a_n \to 0$。

类似地,我们记 $a_n = O(b_n)$,若 $a_n/b_n = O(1)$,或者等价地,对某些常数 C 及所有 n 足够大的值,有 $a_n \leqslant Cb_n$。

我们记 $a_n = o(b_n)$,若当$\to \infty$ 时,$(a_n/b_n) \to 0$。

在下面的例子和先前的章节,当我们说对所有 n 一些东西成立时,我们的意思是对所有 $n \in \mathbb{N}_+ = \{1,2,\cdots\}$,其中我们使用$\mathbb{N}_+$来表示正整数的集合。

例 A.12

(i) 若 $a_n = n/(n+1)$,那么 $a_n = O(1)$,因为对所有 n,$a_n \leqslant 1$。

685

(ii) 若 $a_n = 10/(n+1)$,那么 $a_n = o(1)$,因为当 $n \to \infty$ 时,$a_n \to 0$。

(iii) 若 $a_n = n+5$,$b_n = n$,那么 $a_n = O(b_n)$,因为对 $n \geqslant 5$,$a_n \leqslant 2b_n$,或者对所有 n, $a_n \leqslant 6b_n$。

(iv) 若 $a_n = 1/n$,$b_n = 1/n^2$,那么 $b_n = o(a_n)$,因为 $b_n/a_n = (1/n) \to 0$。

定义 A.48(概率阶数:大 $O_p(\cdot)$和小 $o_p(\cdot)$) 一个实数(可能是向量值)随机变量序列$\{\mathcal{X}_n\}_{n=1}^{\infty}$被称为概率有界,若对每一个 $\varepsilon > 0$,存在一个常数 M 和一个正整数 N(通常 $M = M_\varepsilon$,$N = N_\varepsilon$)使得对所有 $n \geqslant N$,有

$$P[\,\|\mathcal{X}_n\| > M] \leqslant \varepsilon \tag{A.16}$$

即我们说$\mathcal{X}_n$依概率有界,若对任何任意小的正数 ε,我们总是可以找到一个正的常数 M 使得比 M 大的$\mathcal{X}_n$的绝对值(或范数)的概率小于 ε。

很明显，若$\mathcal{X}_n = O(1)$（有界），那么$\mathcal{X}_n = O_p(1)$；然而，反过来就不成立了。令$\{\mathcal{X}_n\}_{n=1}^{\infty}$表示从一个独立同分布$N(0,1)$中随机抽取，那么$\mathcal{X}_n \neq O(1)$，但是$\mathcal{X}_n = O_p(1)$。事实上，任何一个具有良好定义的累积分布函数的随机变量是一个$O_p(1)$变量（见习题 A.5）。

方程（A.16）可以等价地写为对所有$n \geq N$，有

$$P[\ \| \mathcal{X}_n \| \leq M] > 1 - \varepsilon \tag{A.17}$$

我们记$\mathcal{X}_n = O_p(1)$来表示$\mathcal{X}_n$依概率有界。

我们记$\mathcal{X}_n = o_p(1)$，若$\mathcal{X}_n \xrightarrow{p} 0$。

类似地，我们记$\mathcal{X}_n = O_p(\mathcal{Y}_n)$，若$(\mathcal{X}_n/\mathcal{Y}_n) = O_p(1)$，$\mathcal{X}_n = o_p(\mathcal{Y}_n)$若$(\mathcal{X}_n/\mathcal{Y}_n) = o_p(1)$。

注意若$\mathcal{X}_n = o_p(1)$，则$\mathcal{X}_n = O_p(1)$必然成立。但是，当$\mathcal{X}_n = O_p(1)$时，$\mathcal{X}_n$可能不是$o_p(1)$。

如同我们通过计算一个随机变量的均方误差（或r次均值）来获得它的概率极限的情况，我们也可以通过计算它二次矩的阶（或r次均值的阶）来判断一个随机变量的概率阶数，如同接下来定理表明的。

定理 A.7　令$\{\mathcal{X}_n\}_{n=1}^{\infty}$是实数（可能是向量值）随机变量的一个序列，令$a_n$和$b_n$是一些非随机、非负的序列，那么　686

（i）若$E\| \mathcal{X}_n \| = O(a_n)$，那么$\mathcal{X}_n = O_p(a_n)$。

（ii）若$E[\ \| \mathcal{X}_n \|^2] = O(b_n)$，那么$\mathcal{X}_n = O_p(b_n^{1/2})$。

（i）的证明：由$E\| \mathcal{X}_n \| = O(a_n)$，我们知道对某些$m_0 > 0$，有

$$E\| \mathcal{X}_n/a_n \| \leq M_0$$

对任何$\varepsilon > 0$，选择$M = M_0/\varepsilon$（一个有限正常数）。那么根据 Markov 不等式（见式（A.24）），我们有

$$P(\ \| \mathcal{X}_n/a_n \| > M) \leq E\| \mathcal{X}_n/a_n \|/M \leq \varepsilon$$

其意味着$\| \mathcal{X}_n/a_n \| = O_p(1)$或者$\| \mathcal{X}_n \| = O_p(a_n)$。 □

（ii）的证明：（ii）的证明按照类似的方式且被留作习题（见习题 A.6）。

例 A.13

（i）对于随机变量$\{\mathcal{X}_n\}_{n=1}^{\infty}$的任何序列，若对所有$n$有$E|\mathcal{X}_n| \leq C < \infty$，或者若对所有$n$有$E[\ \| \mathcal{X}_n \|^2] \leq C < \infty$，那么$\mathcal{X}_n = O_p(1)$（这直接从定理 A.7 得到）。

（ii）令$\mathcal{X}_n$是具有$E(\mathcal{X}_n) = o(1)$和$\mathrm{var}(\mathcal{X}_n) = o(1)$的一个随机变量序列，那么$\mathcal{X}_n = o_p(1)$。这可以以下事实以及定理 A.7（ii）中得到：

$$E[\ \| \mathcal{X}_n \|^2] = \mathrm{tr}\{E[\mathcal{X}_n \mathcal{X}_n']\}$$

$$= E(\mathcal{X}_n')E(\mathcal{X}_n) + \mathrm{tr}\{\mathrm{var}(\mathcal{X}_n)\} = o(1)$$

定义 A.49(随机等连续) 令 $J_n(b)$ 表示一个由 $b \in B$ 标识的随机过程,其中 B 是 $\mathbb{R}^q$ 的一个紧(闭的和有界的)子集,我们说 $J_n(b)$ 对 $b \in B$ 是随机等连续的,若对所有 $\varepsilon > 0$,我们有

$$\lim_{\delta \to 0} \limsup_{n \to \infty} P\left[\sup_{b, b' \in B, \rho(b, b') < \delta} |J_n(b) - J_n(b')| > \varepsilon\right] = 0 \quad (A.18)$$

其中 $\rho(\cdot, \cdot)$ 是一个度量,比如 $\rho(b, b') = \|b - b'\|$,度量通过欧几里得范数来定义。

随机等连续的一个简单的充分条件是通过如下定理给出的(假定度量 $\rho(x, y) = \|x - y\|$,欧几里得范数)。

687

定理 A.8 令 $J_n(b)$ 表示一个由 $b \in B$ 标识的随机过程,其中 B 是 $\mathbb{R}^q$ 的一个紧子集。若对所有 $b, b' \in B$,我们有

$$E[|J_n(b') - J_n(b)|^{\alpha}] \leqslant C\|b - b'\|^{\gamma} \quad (A.19)$$

对某些 $\alpha > 0, \gamma > 1, C$ 是一个有限的正常数,那么 $J_n(b)$ 对于 $b \in B$ 是随机等连续的。

证明:可以证明式(A.19)隐含着如下条件(见习题 A.12):

对所有 $b, b', b'' \in B$,

$$E[|J_n(b) - J_n(b')|^{\beta}|J_n(b) - J_n(b'')|^{\beta}] \leqslant C\|b'' - b'\|^{\gamma} \quad (A.20)$$

对某些 $\beta > 0, \gamma > 1$,其中 C 是一个正常数。

根据 Billingsley(1968, p.128)的定理 15.6,我们知道式(A.20)是一个保证 $J_n(b)$ 对于 $b \in B$ 是随机等连续的充分条件。因此,式(A.19)意味着 $J_n(b)$ 在 $b \in B$ 上是随机等连续的。

在实践中,当 $J_n(\cdot)$ 有零均值 m,在应用式(A.19)时可以很方便地选择 $\alpha = 2, \gamma = 2$。

定义 A.50(弱收敛) 一个随机元素 $\mathcal{Z}_n(\cdot)$ 的序列(比如 Banach 值或者 Hilbert 值)弱收敛于 $\mathcal{Z}(\cdot)$,若对所有实值有界连续函数 $g(\cdot)$,有 $E[g(\mathcal{Z}_n)] \to E[g(\mathcal{Z})]$。

表述"$\mathcal{Z}_n(\cdot)$ 弱收敛于 $\mathcal{Z}(\cdot)$"也被一些作者表述成"$\mathcal{Z}_n(\cdot)$ 依分布收敛于 $\mathcal{Z}(\cdot)$"。我们在第 13 章考虑的一个非平滑检验统计量的形式是

$$\mathcal{Z}_n(x) = n^{-1/2}\sum_{i=1}^{n} u_i \mathbf{1}(X_i \leqslant x)$$

($x \in \mathcal{S}$,$\mathcal{S}$是一个紧集,即一个闭的和有界的集合),其中(X_i, u_i)是满足 $E(u_i|X_i) = 0$ 的独立同分布数据。$\mathcal{Z}_n(x)$ 是由 x 标识的一个随机过程(或者一个随机元素)。若我们通过下式来定义一个 L_2 范数:

$$\{E[\|Z(\cdot)\|_{\nu}^{2}]\}^{1/2}=\left\{E\left[\int_{S}Z(x)^{2}\nu(\mathrm{d}x)\right]\right\}^{1/2}$$

那么$\mathcal{Z}_n(\cdot)$是一个 Hilbert 值随机元素。可以利用在 1.4 节给出的 Hilbert 值中心极限定理来推导$\mathcal{Z}_n(\cdot)$的渐近分布。

定义 A.51(连续映射定理)

(i) 令 X_n 是一个随机变量序列。若 $g(\cdot)$是一个在 Borel 可测集的每一个点都连续的函数,且 $X_n\to x$ 依分布(或者依概率)收敛,那么 $g(X_n)\to g(x)$依分布(或者依概率)收敛。

(ii) 连续映射定理对于随机元素(或者随机过程)也是成立的,其中 $X_n(\cdot)$和 $X(\cdot)$是(Banach 值或者 Hilbert 值)随机元素,并将上面的依分布收敛替换为弱收敛。 688

1.4 不等式、大数定律和中心极限定理

引理 A.2(Khinchin 大数定律) 若 $X_1,\cdots,X_n$ 是独立同分布的观测值,具有均值$\mu<\infty$,那么

$$\bar{X}_n\equiv\frac{1}{n}\sum_{i=1}^{n}X_i\xrightarrow{p}E\left[\frac{1}{n}\sum_{i=1}^{n}X_i\right]=E(X_i)=\mu$$

证明:注意上面的引理没有要求 $\mathrm{var}(X_i)$是有限的。然而,若 $\mathrm{var}(X_i)=\sigma^2<\infty$,那么例子 A.11(ii)提供了引理 A.2 的一个证明。 □

引理 A.3(Lindeberg-Levy 中心极限定理) 若 $X_1,\cdots,X_n$ 是独立同分布的观测值,具有有限的$\mu<\infty$和方差σ^2,那么

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)\right\}\equiv\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(X_i-\mu)\xrightarrow{d}N(0,\sigma^2)$$

证明:参见 Rao(1973, p.127)。 □

引理 A.4(Lindeberg-Feller 中心极限定理) 令 $X_1,\cdots,X_n$ 是独立的观测值,有 $E(X_i)=\mu_i$, $\mathrm{var}(X_i)=\sigma_i^2$ 以及累积分布函数 $F_i(\cdot)$。定义

$$\bar{\sigma}_n=\left[\frac{1}{n}\sum_{i=1}^{n}\sigma_i^2\right]^{1/2}$$

若

$$\lim_{n\to\infty}\{\max_{1\leqslant i\leqslant n}\sigma_i/[\sqrt{n}\,\bar{\sigma}_n]\}=0$$

$$\lim_{n\to\infty}\frac{1}{n\bar{\sigma}_n^2}\sum_{i=1}^{n}\int_{|x-\mu_i|>\varepsilon\bar{\sigma}_n\sqrt{n}}(x-\mu_i)^2\mathrm{d}F_i(x)=0$$

那么

$$\frac{1}{\sqrt{n}\bar{\sigma}_n}\sum_{i=1}^{n}(X_i-\mu_i)\xrightarrow{d}N(0,1)$$

证明:参见 Rao(1973, p.127)。 □

689 **引理 A.5(Liapunov 中心极限定理)** 令$\{Z_{n,i}\}$是独立(双数组)随机变量的一个序列,使得

$$E(Z_{n,i})=\mu_{n,i},\quad \mathrm{var}(Z_{n,i})=\sigma_{n,i}^2$$

对某些$\delta>0$有$E|Z_{n,i}|^{2+\delta}<\infty$。令

$$S_n=\sum_{i=1}^{n}Z_{n,i},\quad \sigma_n^2=\mathrm{var}(S_n)=\sum_{i=1}^{n}\sigma_{n,i}^2$$

如果

$$\lim_{n\to\infty}\frac{1}{\sigma_n^{2+\delta}}\sum_{i=1}^{n}E|(Z_{n,i}-\mu_{n,i})|^{2+\delta}=0,\text{某些 }\delta>0 \tag{A.21}$$

那么

$$\sigma_n^{-1}(S_n-E(S_n))=\sigma_n^{-1}\sum_{i=1}^{n}[Z_{n,i}-E(Z_{n,i})]\xrightarrow{d}N(0,1) \tag{A.22}$$

引理 A.6(Gramer-Wold 定理) 随机向量$\{X_n\}$的序列$X_n=\{X_{1n},\cdots,X_{qn}\}\in\mathbb{R}^q$依分布收敛于具有累积分布函数$F(\cdot)$的随机向量$X$,若对任何实向量常数$\lambda=(\lambda_1,\cdots,\lambda_q)$,有$\lambda'X_n\xrightarrow{d}\lambda'X$。

通过研究随机变量$\lambda'X_n$的一个标量序列,Gramer-Wold 定理可以被用来推导$X_n\in\mathbb{R}^q$的一个随机向量序列的渐近分布,如以下推论所证明的。

推论 A.1 若对于所有向量$\lambda\in\mathbb{R}^q$,

$$\lambda'X_n\xrightarrow{d}\lambda'X\sim N(\lambda'\mu,\lambda'\Omega\lambda)$$

那么$X_n\xrightarrow{d}N(\mu,\Omega)$。

证明:这个直接可以从引理 A.6 得到。 □

引理 A.7(Borel-Cantelli 引理) 令X_n是一个随机变量序列,令a_n是非负数的一个序列。若$\sum_{n=1}^{\infty}P(|X_n|>a_n)$是有限的,那么

$$X_n\leqslant a_n\quad \text{a.s.},\text{或者 }X_n=O(a_n)\quad \text{a.s.}$$

引理 A.8(Markov 不等式) 假定$\phi:\mathbb{R}^q\to\mathbb{R}$,且对所有$x\in\mathbb{R}^q$,有$\phi(x)\geqslant 0$。令$A\subset\mathbb{R}^q$是$\mathbb{R}^q$的一个子集,并定义$\phi_A=\inf_{x\in A}\phi(x)$,那么

$$\phi_A\mathrm{P}(X\in A)\leqslant E[\phi(X)] \tag{A.23}$$

690 证明:

$$E[\phi(X)]=\int f(x)\phi(x)\mathrm{d}x\geqslant\int_A f(x)\phi(x)\mathrm{d}x\geqslant\phi_A\int_A f(x)\mathrm{d}x$$

$$= \phi_A \mathrm{P}(X \in A) \qquad \square$$

注意当 $x \in \mathbb{R}$ 时，有 $\phi(x) = |x|^k$ 以及 $A = \{x: |x| \geqslant \alpha\}(\alpha > 0)$，那么我们有 $\phi_A = \alpha^k$，$P(X \in A) = P(|X| \geqslant \alpha)$ 以及 $E[\phi(X)] = E[|X|^k]$。那么式(A.23)变成

$$P(|X| \geqslant \alpha) \leqslant E[|X|^k]/\alpha^k \tag{A.24}$$

很多教材将式(A.24)称为“Markov 不等式”。

注意当我们选择 $\phi(x) = \exp(ax)$ 以及 $A = \{x: x > c\}(a > 0, c > 0)$ 时，那么

$$\phi_A = \exp(ax), \quad P(X \in A) = P(X > c)$$

以及

$$E[\phi(X)] = E[\exp(aX)]$$

根据 Markov 不等式，我们有

$$P(|X| \geqslant c) \leqslant E[\exp(aX)]/\exp(ac) \tag{A.25}$$

这是在定理 1.4 的证明用到的式(1.55)。

引理 A.9(Chebychev 不等式)　令 $g(x)$ 是 $\mathbb{R}$ 上的一个正的 Borel 可测函数，其在 $(0, \infty)$ 是单调递增的，$g(x) = g(-x)$。那么对于在 $\mathbb{R}$ 上的每一个随机变量 X 以及 $\varepsilon > 0$，我们有

$$P[|X| > \varepsilon] \leqslant E[g(X)]/g(\varepsilon)$$

当 $g(X) = |X|^k (k > 0)$ 时，我们有 $P[|X| > \varepsilon] \leqslant E(|X|^k)/\varepsilon^k$。

引理 A.10(Hölder 不等式)　令 X_1、X_2 是两个随机变量。那么对 $p > 1$ 以及 $(1/p) + (1/q) = 1$，有

$$E[|X_1 X_2|] \leqslant \{E|X_1|^p\}^{1/p} \{E|X_2|^q\}^{1/q}$$

对 $p = q = 2$，我们得到了著名的 Cauchy(或 Cauchy-Schwarz)不等式。

引理 A.11(迭代期望定律)　令 X 和 Y 是两个随机变量，令 $E(XY)$ 是有限的(且良好定义的)，那么

$$E(XY) = E[XE(Y|X)]$$

引理 A.12(r 次均值收敛)　若(i) $X_n \xrightarrow{p} X$ 或者 $X_n \xrightarrow{d} X$，(ii) $\{X_n^r\}$ 是一致可积的，那么　691

$$E[X_n^r] \to E[X^r], E|X_n|^r \to E|X|^r$$

使得 X_n^r 一致可积的充分条件是对某些 $\varepsilon > 0$，有

$$\sup_n E|X_n|^{r+\varepsilon} < \infty$$

或者，对所有(足够大的)正整数 n，对某些 $\varepsilon > 0$ 以及某些正常数 C，有 $E[|X_n|^{r+\varepsilon}] < C$。

这里 $\sup_n A_n$ 定义如下。若 $A = \sup_n A_n$，那么对所有足够大的 n，有 $A_n \leqslant A$，并

且对任何 $\varepsilon>0$，存在一个 n_0，使得 $A_{n0}>A-\varepsilon$。也就是说，$A=\sup_n A_n$ 是关于序列 A_n 的一个紧上界。

引理 A.13（占优收敛定理（i）） 令 g_n 是定义在集合 S 上的一个可测函数序列，且假定对所有 $x\in S$，有 $|g_n(x)|\leqslant m(x)$，且 $\int_S m(x)\,\mathrm{d}x$ 是有限的。若对所有 $x\in S$，有 $\lim_{n\to\infty}g_n(x)=g(x)$，那么

$$\lim_{n\to\infty}\int_S g_n(x)\,\mathrm{d}x=\int_S[\lim_{n\to\infty}g_n(x)]\,\mathrm{d}x=\int_S g(x)\,\mathrm{d}x \tag{A.26}$$

占优收敛定理给出了可以交换极限和积分运算顺序的条件。若 S 是一个有界的集合（即有有限测度），那么可以通过一个有限常数 C 来替换有界函数 $m(x)$。方程（A.26）在这种条件下仍然成立，且这种情况经常被称为"有界收敛定理"。

引理 A.14（占优收敛定理（ii）） 有一个占优收敛定理的随机版本。若 $X_n\xrightarrow{p}X$ 以及 $|X_n|\leqslant Y$ a.s.，其中 $E(Y^r)<\infty$，那么 $E[|X_n-X|^r]\to 0$ 且 $E(X_n^r)\to E(X^r)$。

证明：参见 Rao（1973）。 □

引理 A.15（可变核的 U 统计量 H 分解） 此处我们通过式（A.27）提供对于二阶 U 统计量 H 分解的一个直观解释：

$$\mathcal{U}_n=\frac{2}{n(n-2)}\sum\sum_{1\leqslant i<j\leqslant n}H_n(X_i,X_j) \tag{A.27}$$

692 其中 $H_n(\cdot,\cdot)$ 是一个对称函数。令

$$H_{1n}(X_i)=E[H_n(X_i,X_j)\mid X_i]$$

那么，H 分解涉及以与微分次序不相关的项的形式重写 $\mathcal{U}_n$，即

$$\begin{aligned}\mathcal{U}_n=&E[H_n(X_i,X_j)]+\frac{2}{n}\sum_i\{H_{1n}(X_i)-E[H_{1n}(X_i)]\}\\&+\frac{2}{n(n-1)}\sum\sum_{1\leqslant i<j\leqslant n}\{H_n(X_i,X_j)-H_{1n}(X_i)\\&-H_{1n}(X_j)+E[H_n(X_i,X_j)]\}\end{aligned} \tag{A.28}$$

若 $E[H_n^2(X_i,X_j)]=O(1)$，那么很容易看出式（A.28）的三个项分别是 $O_p(1)$、$O_p(n^{-1/2})$ 和 $O_p(n^{-1})$。此外，三个项是彼此不相关的。在下面我们 H 分解的应用中，通常 $E[H_n(X_i,X_j)]=O(a_n)$（比如 $a_n=O((h^2+\lambda)^2)$），分解中的第二项是 $O_p(n^{-1/2}a_n)$ 阶的，第三项甚至是更小的阶。我们也使用一个三阶 U 统计的 H 分解，而 Lee（1990，1.6 节）提供了关于一个一般的 k 阶 U 统计 H 分解的详细结论。对于具有可变核的 U 统计量，参见 Powell 等（1989）。

引理 A.16(退化 U 统计量的中心极限定理)　考虑一个二阶 U 统计量,被给定为

$$\mathcal{U}_n = \begin{bmatrix} n \\ 2 \end{bmatrix}^{-1} \sum_{i=1}^{n} \sum_{j>i}^{n} H_n(X_i, X_j)$$

其中 X_i 是独立同分布,H_n 是对称的、位于中心的($E[H_n(X_1,X_2)]=0$),且几乎确定退化的($E[H_n(X_1,X_2)|X_1]=0$),并且 $\sigma_n^2 = E[H_n^2((X_1,X_2)] < \infty$。定义 $G(X_1,X_2)=E[H_n(X_1,X_3)H_n(X_2,X_3)|X_1,X_2]$,那么若当 $n > \infty$ 时,有

$$\frac{E[G_n^2(X_1,X_2)] + n^{-1}E[H_n^4((X_1,X_2)]}{\{E[H_n^2((X_1,X_2)]\}^2} \to 0$$

则 $\mathcal{U}_n / \sqrt{2\sigma_n^2} \xrightarrow{d} N(0,1)$。

证明:参见 Hall(1984)定理 1。　□

693

引理 A.17　令 U_n 是一个二阶 U 统计量,

$$U_n = \begin{bmatrix} n \\ 2 \end{bmatrix}^{-1} \sum_{i=1}^{n} \sum_{j>i}^{n} H_n(Z_i, Z_j)$$

定义

$$r_n(Z_i) = E[H_n \mid (Z_i, Z_j) \mid Z_i]$$

$$\bar{r}_n = E[r_n(Z_i)] = E[H_n(Z_i, Z_j)]$$

$$\overline{U}_n = \bar{r}_n + \frac{2}{n}\sum_{i=1}^{n}[r_n(Z_i) - \bar{r}_n]$$

若 $E[|H_n(Z_i,Z_j)|^2] = o(n)$,那么

(i) $U_n = \bar{r}_n + o_p(1)$。

(ii) $\sqrt{n}(U_n - \overline{U}_n) = o_p(1)$。

证明:参见 Powell 等(1989)的引理 3.1。　□

引理 A.18　对于一个 k 阶退化 U 统计量,有

$$U_n = \begin{bmatrix} n \\ k \end{bmatrix}^{-1} \sum_{(n,k)} H_n(X_{i1}, \cdots, X_{ik})$$

其中 H_n 是一个依赖于 n 的对称(可交换)函数,X_i 是独立同分布随机向量,$\sum_{(n,k)}$ 扩展到$\{1,\cdots,n\}$ 中 $1 \leqslant i_1 < \cdots < i_k \leqslant n$ 的所有组合。假定 H_n 是位于中心的、退化的($E[H_n(X_1,\cdots,X_k)|X_1]=0$a. s.),且定义,对 $c=1,\cdots,k$,有

$$H_{nc}(x_1,\cdots,x_c) = E[H_n(X_1,\cdots,X_k) \mid X_1 = x_1,\cdots,X_c = x_c]$$

它们的方差是 $\sigma_{n,c}^2 = \mathrm{var}[H_{nc}(\cdot)]$。定义

$$G_n(X_1,X_2) = E[H_{nc}(X_1,X_3)H_{n2}(X_2,X_3) \mid X_1,X_2]$$

若对每一个 n 有 $E[H_n^2((X_1,\cdots,X_k)]<\infty$，对 $c=3,\cdots,k$（当 $k\geqslant 3$ 时）有 $\sigma_{nc}^2/\sigma_{n2}^2=o(n^{(c-2)})$，且当 $n\to\infty$ 时，

$$\frac{E[G_n^2(X_1,X_2)]+n^{-1}E[H_{n2}^4(X_1,X_2)]}{\{E[H_{n2}^2(X_1,X_2)]\}^2}\to 0 \qquad (A.29)$$

那么 nU_n 是渐近正态的，具有零均值和方差 $k^2(k-1)^2\sigma_{n2}^2/2$。

Fan 和 Li(1996)证明了以上结论。当 $k=2$ 时，它简化为 Hall(1984)的定理 1。

令 ν 表示一个 L_2 范数，令 $\mathcal{S}$ 是一个具有 $\nu(\mathcal{S})<\infty$ 的 Hilbert 值随机元素的支撑。若 $\mathcal{S}$ 是 $\mathbb{R}^q$ 的一个有界子集，那么可以选择 $\nu(\cdot)$ 作为 $\mathcal{S}$ 上的 Lebesgue 测度。

694

引理 A.19(Hilbert 值中心极限定理) 令 $Z_1(\cdot),\cdots,Z_n(\cdot)$ 是 $\mathcal{L}_2(\mathcal{S},\nu)$ 上 Hilbert 值独立同分布零均值随机元素，使得

$$E[\|Z_i(\cdot)\|_\nu^2]=E\left[\int Z_i^2(x)\nu(\mathrm{d}x)\right]<\infty$$

那么

$$\mathcal{Z}_n(\cdot)\stackrel{\text{def}}{=}n^{-1/2}\sum_{i=1}^n Z_i(\cdot)$$

弱收敛于(比如 $\mathcal{Z}_\infty(\cdot)$)一个协方差函数为 $\Omega(x,x')=E[Z_i(x)Z_i(x')]$ 的零均值高斯过程。

证明：参见 Politis 和 Romano(1994)的定理 2.1，或者参见 van der Vaart 和 Wellner(1996，习题 1.8.5，p.50)。 □

注意对于弱收敛通常需要检验两个条件：(i) 有限维收敛，即 $(\mathcal{Z}_n(x_1),\cdots,\mathcal{Z}_n(x_m))$ 弱收敛于 $(\mathcal{Z}_n(x_1),\cdots,\mathcal{Z}_n(x_m))$；(ii) 过程(由·标识) $\mathcal{Z}_n(\cdot)$ 是紧的。[①] 第(i)点从 Lindeberg-Levi 中心极限定理和 Gramer-Wold 策略得到，而(ii)从 $E[\|Z_i(\cdot)\|_\nu^2]<\infty$ 得到，其是保证过程 $n^{-1/2}\sum_{i=1}^n Z_i(\cdot)$ 是紧的充分条件(在 L_2 范数 $\|\cdot\|_\nu$ 下)。

若不使用我们在第 13 章讨论过的 CM 类型统计量，可以选择基于 sup 范数的统计量，比如

$$I_n(x)\stackrel{\text{def}}{=}\sup_x\left|n^{-1/2}\sum_{i=1}^n\mathbf{1}(X_i\leqslant x)\right|$$

那么就使用了一个 Banach 值随机元素(即不是 Hilbert 值，因为 sup 范数不能够

① 对于一个随机元素序列是紧的定义参见 Billingsley(1968，p.40)。

从一个内积得出)。为了证实一个 Banach 值随机元素序列是紧的,可以参见 Billingsley(1968)和 Pollard(1984);也可参见 Stute(1997),他讨论了具有$I_n(\cdot)$形式的一个特定统计量。

1.5 习 题

习题 A.1

(i) 证明所有有理数的集合是一个 Borel 集。

(ii) 证明所有无理数的集合是一个 Borel 集。

提示:

(i) 注意有理数的集合是可数的,即它可以被表示为 $\sum_{i=1}^{\infty}\{a_i\}$,其中每一个 a_i 是一个不同的有理数(单点)。

(ii) 无理数是有理数的补集。

695

习题 A.2 令 $g(x)$ 被定义为

$$g(x) = \begin{cases} 1, & \text{若 } x \text{ 是无理数} \\ 0, & \text{若 } x \text{ 是有理数} \end{cases}$$

这被称为 Dirichlet 函数。证明 Dirichlet 函数是一个 Borel 可测函数。

提示:考虑形如$\{x:g(x)\leqslant a\}$的集合,对任何常数 $a<0$, $a=0$, $0<a<1$, $a=1$,以及 $1<a<+\infty$。注意空集、有理数集、无理数集和$\mathbb{R}$都是 Borel 集。

习题 A.3 令 A 表示在$[0,1]$内的所有有理数,B 表示在$[0,1]$内的所有无理数。A 和 B 的 Lebesgue 测度是什么?

提示:$[0,1]=A\cup B$, $[0,1]$有 1 的 Lebesgue 测度,并且

$$1 = m([0,1]) = m(A) + m(B)$$

因为 $A\cap B=\phi$。

习题 A.4 在更强的条件 $E[(X_n-X)^2]\to 0$ 和 $E[(Y_n-c)^2]\to 0$ 下,证明定理 A.6(i)

习题 A.5 证明具有一个良好定义的累积分布函数即 $F(x)$ 的任何随机变量是一个 $O_p(1)$ 变量。

提示:一个良好定义的累积分布函数具有如下性质:(i) $\lim_{x\to\infty}F(x)=1$,(ii) $\lim_{x\to-\infty}F(x)=0$,(iii) $F(x)$是一个非递减的函数,(iv) $F(x)$是右连续的。在这个练习中仅仅只需使用(i) 和(ii)。

习题 A.6

(i) 令 a_n 和 b_n 是两个任意正的序列,证明

$$O(\sqrt{a_n+b_n}) = O(\sqrt{a_n}) + O(\sqrt{b_n})$$

(ii) 证明定理 A.7(ii)。

提示:(i) 利用

$$\sqrt{a_n+b_n} \leqslant \sqrt{a_n} + \sqrt{b_n} \leqslant 2\sqrt{a_n+b_n}$$

(通过平方每一项可以很容易地看到不等式成立。)(ii) 利用 Chebychev 不等式。

习题 A.7 证明练习 A.2 定义的 Dirichlet 函数是 Lebesgue 积分且在 $x \in [0,1]$上计算它的 Lebesgue 积分。

696

提示:Dirichlet 函数是一个由 $g(x) = (1)\mathbf{1}(A) + (0)\mathbf{1}(B)$给出的简单函数,其中 A 是无理数的集合,B 是有理数的集合。

注意 Dirichlet 函数很明显不是 Riemann 积分。

习题 A.8 计算 Lebesgue 积分(i) $\int_0^1 x^2\mu(\mathrm{d}x)$ 和(ii) $\int_0^{\pi} \cos(x)\mu(\mathrm{d}x)$,其中 μ 是 Lebesgue 测度。

提示:利用定理 A.1 的结论。

习题 A.9 令$\{e_j\}$是 Hilbert 空间的一个正交基。证明 $c_j = (e_j, g)$,其中 c_j 在式(A.7)中被定义。

提示:考虑(g, e_l),用式(A.7)的右边来替换 $g(\cdot)$,并且利用基础函数的正交性质。

习题 A.10

证明:$\{1/\sqrt{2\pi}, \sin(\pi x)/\pi, \cos(\pi x)/\pi, \sin(2\pi x)/\pi, \cos(2\pi x)/\pi, \cdots, \sin(n\pi x)/\sqrt{\pi}, \cos(n\pi x)/\sqrt{\pi}, \cdots\}$是关于 $x \in [-\pi, \pi]$的一个正交序列。

提示:对任何正整数 k、l,有

$$\int_{-\pi}^{\pi} \sin(kx)\,\mathrm{d}x = \int_{-\pi}^{\pi} \cos(lx)\,\mathrm{d}x = 0$$

$$\sin(kx)\cos(lx) = [\sin((k+l)x) + \sin((k-l)x)]/2$$

$$\sin(kx)\sin(lx) = [\cos((k-l)x) - \cos((k+l)x)]/2$$

$$\cos(kx)\cos(lx) = [\cos((k-l)x) + \cos((k+l)x)]/2$$

$$\cos^2(kx) = [1 + \cos(2kx)]/2$$

$$\sin^2(kx) = 1 - \cos^2(kx)$$

习题 A.11 证明$\int_{-\infty}^{\infty} \delta(x-a)g(x)\,\mathrm{d}x = g(a)$,其中 $\delta(\cdot)$是 Dirac delta 函数。

提示:记 $g(x) = g(a) + [g(x) - g(a)]$,并且注意$[g(x) - g(a)]\delta(x-a)$

$=0$。

习题 A.12 令 $J_n(b)$ 是定理 A.8 给出的一个随机过程。证明式(A.19)暗含着如下结论：

对所有 $b, b' \in B$，b'' 在 b 和 b' 的线段间，

$$E[|J_n(b) - J_n(b'')|^{\beta} |J_n(b') - J_n(b'')|^{\beta}] \leqslant C \| b'' - b' \|^{\gamma} \quad (A.30)$$

对某些 $\beta > 0$，$\gamma > 1$，C 是一个正的常数。

提示：利用 $|ab| \leqslant (1/2)(a^2 + b^2)$，选择 $\alpha = 2\beta$，且注意

$$\max\{ \| b' - b'' \|, \| b - b'' \| \} \leqslant \| b' - b \|$$

主题索引

译 后 记

我们有幸承担这本专著的翻译工作。该专著涉及该领域的方方面面，翻译难度比较大。因为该专著涉及很多该领域的前沿研究成果，所以，翻译这本专著对我们的帮助也比较大。该专著的初译工作由孟祥财（负责第 1 至第 3 章）、李昕（负责第 4 至第 7 章）、郭光远（负责第 8 至第 11 章）、褚冬琳（负责第 12 至第 14 章）、仇静（负责第 15 至第 18 章）和罗彩琴（负责第 19 至第 20 章及附录）完成。对翻译初稿的文字校对工作由博士生吴相波完成，公式的校对工作由博士生吴继贵完成。最终的校对工作由叶阿忠完成。感谢责任编辑郝小楠对译稿认真负责的编辑和校对，使译稿增色不少。由于我们的水平和能力有限，虽然已经尽了我们的努力，但译文中的错误与纰漏在所难免，请读者和专家指正，译者的联系方式为 E-mail：ye2004@fzu.edu.cn。

叶阿忠

于福州大学新校区

2013 年 7 月